Does Relative Performance Information Lower Group Morale?

Lea Heursen
Department of Economics
Humboldt University of Berlin

This version: July 2019*

Abstract

In many organizations, productivity relies not just on individual effort but also on group morale, that is, the willingness of co-workers to help each other perform better at work. Relative performance evaluations (RPE) are known to increase individual work morale but may negatively affect group morale because they create a sense of competition among members of a reference group. In a novel experiment, I vary whether or not members of a reference group obtain relative performance information on a task that is relevant for their social image or self-image, a general knowledge test, and measure how this affects the subsequent willingness to help the productivity of others by sharing knowledge with them at a personal cost. I find that, compared to a baseline with no relative performance information and fixed piece-rates, RPE cause members of a reference group to compete as intensely as under relative pay and also increases the perceived social distance between them. Yet, I show that even after a performance competition, individuals are willing to help the productivity of others in the group. These findings advance our understanding of how relative concerns among co-workers affect the way they work together.

Keywords: relative performance information, rank feedback, social incentives, on-the-job help, group productivity, social and self-image, experiment

*I am grateful to Roberto Weber, in particular, and Ernst Fehr for their support in the development of this project. I thank Björn Bartling, Yves Breitmoser, Dirk Engelmann, Christine Exley, Eva Ranehill, Ernesto Reuben, Karl Schlag, Julien Senn, Anja Schöttner, Sigrid Suetens, Vanessa Valero, Georg Weizsäcker and Florian Zimmermann for their comments and feedback. This paper also benefited from discussions with the audience in numerous conference and seminar presentations. Financial support from the Excellence Foundation Zurich and the Zurich Graduate School of Economics through their Director's Grant program is gratefully acknowledged.

1. Introduction

Jobs require more and more that employees collaborate, for example, in problem-solving teams (Lazear and Shaw 2007; Deming 2017). Therefore, to understand productivity in organizations, it is important to understand what makes groups of co-workers productive. Social incentives that encourage social comparisons in performance, for example by providing members of a reference group with relative performance feedback, are known to increase the productivity of individuals (Blanes i Vidal and Nossol 2011; Kuhnen and Tymula 2012; Tran and Zeckhauser 2012; Gill et al. 2018). Much less attention has been paid to how these types of social incentives affect other behavior in workgroups. Relative performance feedback seems to encourage cheating and sabotage in reference groups by group members who want to artificially inflate their relative performances (Charness, Masclet and Villeval 2014). How do relative performance evaluations affect how well people work together?

One important factor of how well people work together is *group morale* that I define as *the willingness of co-workers to help each other perform better at work*.[2] High group morale can increase productivity, when co-workers are willing to invest in the performance of others by helping them to perform better on their jobs. Very little is known empirically about the factors that encourage or discourage group morale.

In this study, I investigate how one factor, social information in the form of relative performance evaluations, affects co-workers' willingness to help each other perform better on their jobs. Some of the largest companies in the world, such as Amazon[3] or Yahoo[4], evaluate their employees' relative performances by ranking them. These are also examples of companies in which a lot of the production is done by co-workers who work together. Schools and universities are another class of organizations in which relative performance evaluations are common, for example, when students are "graded on a curve". Given its prevalence in organizations with group production, it is important to understand how relative performance evaluations affect how well people work together. Microsoft is an example of a firm where relative performance feedback appears to have had unintended consequences in workgroups. In November 2013, the company abolished a relative performance evaluation scheme under which employees within organizational units were ranked according to their performances. This ranking system

[2] I draw on this term in reference to individual work morale that describes the willingness to exert more effort in order to perform better individually. I use the term group morale in a narrow sense, to describe the willingness of individuals to help the performance of others, following how economists have integrated the concept of individual (work) morale in their work that also emphasizes the motivation to exert more effort in order to perform better individually.

[3] "Inside Amazon: Wrestling Big Ideas in a Bruising Workplace", The New York Times, accessed April 9, 2018 https://www.nytimes.com/2015/08/16/technology/inside-amazon-wrestling-big-ideas-in-a-bruising-workplace.html

[4] Marissa Mayer famously introduced a relative performance ranking system when she joined Yahoo in 2012, see, e.g., "What Marissa Mayer Got Wrong (and Right) About Stack Ranking Employees", Harvard Business Review, accessed April 9, 2018 https://hbr.org/2015/01/what-marissa-mayer-got-wrong-and-right-about-stack-ranking-employees

was partly blamed for a "lost decade" at the company[5] and highly unpopular with employees. "It leads to employees focusing on competing with each other rather than competing with other companies", as one Microsoft employee describes it. An important rationale for the abolition was to "promote new levels of teamwork" and "put more emphasis on teamwork and collaboration" (Microsoft Humam Ressource Chief).[6]

In this paper, I use a laboratory experiment to investigate whether relative performance evaluations lower group morale. The study is designed to test the following mechanism: relative performance information could cause members of a reference group to compete against one another and this may impact on how well they work together. I construct a design in which I can compare help behavior in reference groups whose members received relative performance feedback to help behavior in reference groups whose members were not evaluated against each other.

The experimental laboratory offers two important advantages over observational data sets. Firstly, in my experimental environment, I can isolate the effect that a competitive environment in and of itself has on the willingness to help co-workers perform better. Career concerns can offer a strategic motive to not invest in the productivity of those who compete for promotions. They are hard to disentangle from relative concerns *per se,* under which co-workers simply want to maintain outcome differences. In the experiment, strategic motives are completely shut down since helping others does not affect the probability of obtaining a high or low relative performance rank. Secondly, I can pin down the effect of relative performance feedback because I can precisely control a major confounding factor present in data from companies such as Microsoft: high relative performances come, at least in the medium run, with a monetary prize such as promotions. Employees may simply compete for the money that comes with high rankings. Alternatively, the relative performance feedback *per se* may be sufficient to spur a competition for non-monetary or image-based rewards.

In my experiment, a total of 282 subjects participate in one of four experimental conditions. I vary experimentally whether members of a reference group receive only absolute feedback on a timed general knowledge test (baseline) or also an evaluation of their relative performances on this test, in an environment with fixed monetary piece-rates for correct answers to questions. In the private rank feedback treatment, a group member observes her performance rank in the group (rank 1, rank 2 or rank 3). In the public rank feedback treatment, I establish common knowledge about performance ranks in the group, by displaying the picture of a group member next to that group member's performance rank. Thus, these two treatments give people the kind of relative performance feedback that has been shown to motivate people to work harder

[5] "Microsoft's Lost Decade", Vanity Fair, accessed 9 April, 2018, https://www.vanityfair.com/news/business/2012/08/microsoft-lost-mojo-steve-ballmer

[6] "Microsoft Axes its Controversial Employee-ranking System", The Verge, accessed on 9. April 2018, https://www.theverge.com/2013/11/12/5094864/microsoft-kills-stack-ranking-internal-structure

individually (see e.g. Blanes i Vidal and Nossol 2011; Tran and Zeckhauser 2012). A control treatment introduces relative pay. A group member observes her performance rank in private and the best performer in a group earns a substantial monetary bonus, in addition to the piece-rate for correct answers.

I use these conditions to test, in a between-subject design, whether relative performance evaluations cause the perception of competition within a reference group and whether this impacts on group members' willingness to help others in their group perform better. With the relative pay control treatment, I explore whether any results change moving from purely image-based rewards to a monetary reward for relative performance.

For this purpose, in the second part of the experiment, I measure group morale in a way that closely resembles the type of help that is important in workgroups: sharing knowledge.

Participants answer more general knowledge questions for a fixed monetary piece-rate. They can invest in the performance of others in their reference group by sharing their answers to this new set of questions at a small personal cost that amounts to 4% of the piece-rate.[7] Sharing answers can improve the performance and earnings of other group members because the computer automatically replaces their incorrect answers with a correct answer that was shared. Group members make their help decisions simultaneously and throughout the experiment nobody observe help behavior of others. The fact that others ignore the help behavior of anybody else means that neither selfish nor prosocial behavior is detectable. Importantly, this design feature rules out that the desire to demonstrate to others a (perceived) advantage in knowledge or a prosocial attitude motivate knowledge-sharing decisions.

I model helping as a pro-social act since it comes at a small cost and benefits others in the group without any direct *monetary* benefit for the person who lends helps. This mirrors the actual cost of helping in firms. Moreover, this type of pro-social help among co-workers should be most sensitive to changes in the intensity of competition among them.

In the third part of the experiment, I measure participants' beliefs about the correctness of their own and their group members' answers to all questions for which help decisions were made, in an incentivized and incentive-compatible way. This data is important for two reasons. First, I can confirm that people share their knowledge in order to make others perform better. Second, I can control for the fact that the rank information of the treatments may change beliefs about the value of own help to others in the group. This may contribute to people helping differently across experimental conditions. I also elicit the sense of competition and social distance in reference groups, as well as beliefs about the expected help by others. With this data,

[7] This piece-rate was calibrated with a pilot study of the baseline condition to ensure that the level of knowledge-sharing in the baseline condition was neither too high (i.e. above 75% or higher) or to low (i.e. below 25%). In a first calibration of the design, the personal cost to sharing answers with others was 10% of the piece-rate under which the level of knowledge-sharing was too low.

I evaluate the impact that relative performance evaluations have on the perception of social relations in reference groups.

I find that both private and public relative performance feedback causes a large and statistically significant increase in the level of perceived competition among group members, compared to the baseline condition with a very low level of perceived competition. I find that the intensity of competition under private or public relative performance evaluations, is, on average, as large as the intensity of competition when group members compete for relative pay. Relative performance evaluations also increase the social distance between members of experimental reference groups.

I then show that even after a performance competition, a substantial share of participants helps others to perform better. In the baseline condition, I observe a whole distribution of levels of help. About 17% of participants never share and 11% share all their answers with others and participants share the answer to 4 out of 10 questions on average. Compared to this baseline behavior, the empirical distributions of help after a performance competition are very similar and I cannot reject the null hypothesis that they are drawn from the same population.

With my data, I can very precisely estimate that the average treatment effect of a performance competition on the subsequent willingness to help the performance of others is close to zero. This is true for the private rank feedback, the public rank feedback and the relative pay treatments.

Binary choice models, in which I predict the probability of sharing answers with others as a function of the treatment status and the expected value of own help to the group, support this main result from parametric and non-parametric hypothesis testing. This analysis shows that participants are much more likely to share answers that they think will improve the performance of their group members. This is consistent with the interpretation that participants use the help technology for the benefit of others in their group.

I estimate that missed opportunities to help—instances in which a group member submitted the wrong answer to a question despite someone in the group submitting the correct answer—amounted to an average efficiency loss of 18% of realized group earnings. This shows that a change in group morale, a decrease or increase, could have real consequences on group productivity in the setting that I study.

I argue that my main finding on group morale under relative performance evaluations is a credible null result. I use standardized effect sizes (d)[8] of the treatment effects from two related studies (Carpenter, Matthews and Schirm 2010; Buser and Dreber 2016) as benchmarks to guide my power analysis that I perform at conventional levels of statistical significance. I find that the chance of a statistical Type-2 error would be as low as 6% if the true standardized effect size of

[8] I define effect size as the standardized difference in the average number of answered shared between the baseline and a treatment condition.

relative performance evaluations on average help in reference groups was close to the average of these two benchmarks (d=0.597).

This study makes several contributions to the literature. It isolates how relative concerns in a non-monetary domain, the performance domain, affect how well members of a reference group work together. In light of the prominence of relative performance information in firms, this is an important gap to fill. Secondly, this study is the first to directly compare the intensity of competition in reference groups under relative performance information and relative pay. Thirdly, this study introduces a way to elicit "real help"[9] behavior in a versatile and easy-to-implement task which models an important dimension of on-the-hob help in organizations. My study offers two insights for performance evaluations in organizations. Relative performance feedback is likely to trigger a comparative and competitive mind-set in reference groups. I find no evidence suggesting that relative concerns among co-workers in an of themselves, caused either by competing for relative performance rankings or relative performance pay, have a substantial negative effect on the willingness of co-workers to help each other perform better at work. My findings thus provide a more positive outlook on relative concerns among co-workers and how they impact on the way co-workers interact and work together, than what related research or the prior views of business insiders[10] may suggest.

2. Related Literature

This study builds on the small but growing empirical literature that finds that relative concerns among co-workers entail behavioral spillovers to workplace behavior. In a laboratory experiment with a real effort task, Carpenter, Matthews and Schirm (2010) find that relative performance pay causes an increase in sabotage among members of a reference group, even if sabotaging others' work has no effect on the likelihood of winning the bonus competition. In a large-scale online experiment, Buser and Dreber (2016) show that competing for relative pay significantly and sizably lowers subsequent contributions in an unrelated Public Goods Game, compared to a baseline condition with a fixed piece-rate.[11] Breza, Kaur, and Shamdasani (2018)

[9] Analogously to how experimental economists use the term "real effort" (see e.g. Charness, Gneezy, and Henderson 2018) the term "real help" is used to indicate that the outcome of the help decision depends on the performances of the person who lends help and of the person who receives help. By contrast, in a "chosen-help" or "stated-help" design, help is modeled as the transfer of money.

[10] In addition to the perception at Microsoft, CEOs of other companies have also voiced concerns about relative evaluations of employees on the ground that they may hinder collaboration. Qualtric's CEO Smith is quoted with the following opinion on relative performance evaluations: "Stack-ranking is fine, says Smith, for evaluating performance in a sales organization, where managers may want to heighten competition. It's less well suited, he says, for evaluating engineers, among whom management may want to create closer collaboration.", "Cooperation "Microsoft: 'Stack-Ranking' Gets Heave-Ho", accessed 9, April , 2018 https://abcnews.go.com/Business/microsoft-abolishes-stack-ranking-employees/story?id=20877556

[11] Independently of this study, Black et al. (2018) investigated a related research question and tested whether relative performance information on a first task spills over to behavior in an unrelated continuous Prisoner's dilemma game that is played with someone outside of the reference group of the first task. They test for a broader spillover to an unrelated cooperative game, similar to Buser and Dreber (2016), and do not find evidence of it. However, the very small sample size of 20-24 participants per experimental condition and the fact that 9 out of 24 participants did not send any money in the dictator game played in the baseline condition--which suggests that the treatment could

investigate in a field experiment how relative pay concerns affect workers' individual work morale and their ability to work together. They find that a shared history of wage disparities, established on the basis of relative performance differences in reference groups, lowers co-workers' ability to produce together in joint production tasks in which it is in the self-interest of coworkers to work together. In all of these studies, it was found that relative pay concerns affect sabotage and cooperation, in the absence of a direct strategic link between cooperation and sabotage and the likelihood of wining the relative performance based prize. My study adds to this literature by investigating, and thereby isolating, whether relative *performance* concerns in and of themselves alter the way members of a reference group work together.

This paper explores a potential cost to relative performance feedback which may come at the expense of a lower group morale. Two related studies reveal different unintended side effects of relative performance feedback in firms. In a field experiment, Bandiera, Barankay and Rasul (2013) show that rank feedback reduces the average productivity in a firm because this information changes how work teams form. In the face of rank feedback about the relative productivities of teams, workers appear to match more on ability than on social ties when teams are re-matched. Charness, Masclet and Villeval (2014) show that public performance rankings cause members of a reference group to spend money in order to artificially increase their own performance, i.e. to cheat, or to lower the output of others in their group, i.e. to sabotage. The empirical evidence suggests that participants sabotage and cheat in order to change the final ranking. The findings of my study draw attention to the fact that more research is needed to understand when social incentives in the form of relative performance feedback do, and when they do not backfire in reference groups.

My two treatments—private and public performance rank feedback on a general knowledge test—build on a large conceptual literature in behavioral economics on people being motivated by self-image (Köszegi 2006) and social image (Bénabou and Tirole 2006; Ellingsen and Johannesson 2008; Besley and Ghatak 2008). Several empirical studies confirm that people like to signal to themselves or to others that they are intelligent (Tran and Zeckhauser 2012, Ewers and Zimmermann 2015). But there are a few studies that find that an audience can affect the desire to signal competence to the observers in surprising ways. McManus and Rao (2015) show that an audience of college peers can change behavior toward actions that are *less* likely to signal high intelligence, even though the trait is privately desirable. A field experiment at a business school provides further evidence that the desirability of a competent or ambitious image seems to depend on the target audience (Bursztyn, Fujiwara, and Pallais 2017), a point that was first made theoretically in a two-audience education signaling model by Austen-Smith and Fryer (2005). I considered these findings in the design of my experiment, attempting to ensure that

potentially only change the behavior of a subset of the 20 participants in the treatment condition--makes it difficult to draw firm conclusions from this study.

performing well and being seen as performing well at the real effort task is desirable for the participants.

My paper contributes to the empirical literature on on-the-job help. While a link between tournaments in firms and worker cooperation has been established theoretically (Lazear 1989), there are only a few papers that investigate empirically the determinants of on on-the-job help. With survey data among employees of a plant, Drago and Garvey (1998) provide correlational evidence that promotion tournaments lower employees' qualitative ratings of their co-workers willingness to provide help (e.g. by sharing tools or machinery). Danilov, Harbring, and Irlenbusch (2014) present findings from a laboratory experiment that models help as the transfer of money and compares help under team-pay and relative performance pay. They find that participants help less as the bonus paid to the best performer increases and help more, *ceteris paribus*, as the return to team-output increases. Unlike in my study, however, the change in helping results directly from the monetary incentives, i.e. the relative pay or the team-pay, that Danilov, Harbring and Irlenbusch paid subjects. The findings oft this present study advance this literature by providing causal evidence on whether relative concerns among co-workers in and of themselves systematically affect the willingness to help.

The rest of the paper is organized as follows, section 3 presents a framework to illustrate the effects under study, in section 4, I outline the design of my experiment in detail. Section 5 shows the results and section 6 discusses them and concludes.

3. Framework

In the following, I present a simple conceptual framework to illustrate how relative performance evaluations can change group morale, that is, the willingness to help others perform better in reference groups. An employee (i) receives a piece-rate b for good performance on a task. The employee either knows how to solve this task ($p_i = 1$) or does not ($p_i = 0$). Her utility is $u_i = bp_i$.

She can lend help $h_i \in \{0,1\}$ to a co-worker (j) in her reference group, for example, by sharing her knowledge on how to solve the task, to improve his performance. Helping comes at a small cost $c > 0$.[12] Her payoff $v(h_i)$ from helping is:

$$v(h_i) = G(a_i, p_i, p_j)u_j(h_i) - ch_i$$

where $G()$ is a group morale effect term. Helping a colleague can increase his performance. I use the notation $u_j(h_i)$ to indicate j's expected monetary payoff as a function of employee i's

[12] For example, there is an opportunity cost of time when she spends some time explaining the solution to her co-worker.

help. Likewise, $u_i(h_j)$ denotes i's expected monetary payoff as a function of the help of colleague j.[13]

Clearly, the employee helps her co-worker whenever

$$\Delta v(h_i) = G(a_i, p_i, p_j)\Delta u_j(h_i) \geq c.$$

The benefit to help is increasing in $G()$ such that, *ceteris paribus*, employees with a higher group morale are more likely to help other colleagues to perform better. Helping is an act that benefits her co-worker, therefore, an employee's group morale is linked to her pro-sociality. I conceptualize it in the following way:

$$G(a_i, p_i, p_{-i}) = a_i - s(p_i, p_j)$$

The employee gives weight $a_i \in [0,1]$ to the utility of her co-worker. There is ample evidence that people differ in their prosocial inclinations. Thus, this simply captures that some colleagues are more inclined to help than others and some may never help (e.g. $a_i = 0$). In addition, relative performance concerns $s(p_i, p_j)$ may change an employee's group morale or, in other words, the value that she assigns to improving the performance and earnings of her colleague. Note that without this second term, the group morale term $G()$ is a simple altruism model. It is common to conceptualize the weight that people give to the utility of others as partly determined by the decision context, for example, by others' behavior (Rabin 1993) or by their pro-social type (Levine 1998) or, as in this framework, by relative performance concerns.

I conceptualize relative performance concerns as a comparison of performance outcomes.[14] For example, an employee compares her own performance rank $r_i(p_i, p_j)$ to the performance rank $r_j(p_i, p_j)$ of a colleague in her reference group:

$$s(p_i, p_j) = \sigma_t f(r_i - r_j)$$

where $f(\cdot): \mathcal{R} \rightarrow [0,1]$.[15] The parameter $\sigma_t \in [0,1]$ represents the extent to which relative performance concerns are prevalent in a reference group and the subscript t stands for treatment. I incorporate this into the employee's group morale term:

$$G(a_i, p_i, p_j) = a_i - \sigma_t f(r_i - r_j) \in [-1,1]$$

This captures the idea that an employee's concern for improving the performance of other co-workers is shaped by her relative performance concerns, when these are salient in her reference

[13] Of course, it may be that $u_j(h_i = 1) = u_j(h_i = 0)$, for example, when group member j knows how to solve the task in which case $u_j(h_i = 1) = u_j(h_i = 0) = b$.

[14] The framework does not explicitly model comparisons of monetary outcomes because, in the experiment, group members did not observed the actual earnings of other group members—since they did not know the absolute performance of others—and because ranks were (weakly) informative about earnings differences in the group.

[15] I do not make any assumptions on the shape of this function $f(\cdot)$ other than bounding the positive or negative impact it can have on the weight that an employee gives to the utility of her co-worker j. In order to detect an effect of relative performance information on group morale in this experiment it just needs to be that $f(x)|_{x\in\{-2,-1,1,2\}} \neq 0$.

groups. Initial empirical work on relative performance concerns and unethical behavior in reference groups suggests that $s(p_i, p_j)$ enters the group morale term negatively (Charness, Masclet and Villeval 2014). So does conceptual work on preferences for status. Note that the group morale term could also be negative, in which case an employee may be willing to sabotage her co-worker.[16]

Treatment Manipulations and Group Morale

I hypothesize that relative performance rankings create a sense of competition in reference groups that activates positional concerns over relative performances. Rustichini (2008, p. 653) summarizes the link between competition and relative concerns in a review article on dominance and competition in the following way: "Humans who participate in a contest with others have strong preferences on relative outcomes, and are ready to translate these preferences into costly choices". This implies the following for the salience of relative performance concerns across experimental conditions:

$$\sigma_{baseline} < \sigma_{privateRF} \leq \sigma_{publicRF}$$

The second weak inequality summarizes my hypothesis about public rank feedback. The provision of public performance ranks may further strengthen relative performance concerns since it explicitly invokes social image in intelligence by establishing common knowledge about every group member's performance rank within the reference group.

As the salience of relative performance concerns increases from $\sigma_{t'}$ to σ_t, group morale decreases. To see this, note that in this case, the benefit of helping another group member perform better decreases by $(\sigma_{t'} - \sigma_t) f(r_i - r_j)$, whenever $f(r_i - r_j) \neq 0$.

Thus, in a between-subject design, I can identify the effect of relative performance concerns on the willingness to help others in the reference group perform better by comparing average help in the baseline condition to average help under private or public rank feedback.[17]

A control treatment introduces relative performance pay. Compared to the baseline condition with fixed piece-rates and no rank information, relative performance concerns should be more salient under relative pay, that is, I hypothesize that $\sigma_{baseline} < \sigma_{relative\ pay}$.[18]

In the next section, I present how I manipulated the salience of relative performance concerns across the four experimental conditions and how I measure group morale. Once the design is introduced, I re-state my hypotheses illustrated with this framework in terms of behavior in the experiment.

---

[16] The possibility for sabotage is ruled out by design in this experiment.

[17] With random assignment to treatment, I can rule out that participants' general inclination to help others, a_i, differs systematically across treatments, that is, with random assignment to treatment it holds that $E[a_i|treatment] = E[a_i]$.

[18] It is not clear whether monetary and non-monetary returns to high relative performance ranks complement or substitute each other. While the salience of relative concerns should not be lower in the relative pay condition, compared to the private rank feedback condition, it need not necessarily be higher.

4. Experiment Design

4.1. Overview

In a first part of the experiment, I induce a sense of competition in reference groups by providing relative performance feedback on a task that people perceive as relevant for their social and self-image: a general knowledge test.[19] In a between-subject design, I vary by experimental condition whether or not group members' performances on the test are evaluated relative to one another. In a control treatment, I introduce relative pay that is based on relative performance on this general knowledge test. I then measure, in a second part of the experiment, how the treatment affects the willingness to help others in the reference group perform better. With this design, I test whether a competitive environment in and of itself has a negative effect on group morale. The final parts of the experiment, 3 and 4, elicit further outcome measures and control variables.

Table 1 summarizes the timeline of the experiment. The following paragraphs provide the details for each part.

Table 1. Timeline of Experiment

Part 1 Performance and Feedback
Stage 1 Measure general knowledge
Stage 2 Timed general knowledge test with varying relative performance feedback and pay (by treatment)
Part 2 Measure Group Morale
Part 3 Measure Beliefs
Part 4 Questionnaire

4.2. Part 1—Performance and Feedback

Part 1 had two stages.

In the first stage of Part 1, every participant was tested on his general knowledge with 10 multiple choice[20] questions. Participants obtained 25 points for each correct answer and these points were converted at a fixed exchange rate to 1.5 CHF (≈1.5 USD) in pay at the end of the experiment. Performance on these first questions measures baseline ability at answering the type of general knowledge questions that are used throughout the study. For this first ability measure, everything was held constant across experimental conditions.

[19] Extensive pre-testing of the task, general knowledge questions, in the same student population as in the main experiment ensured that students wanted to perform well and wanted to be seen as performing well on this task by others.

[20] Each question had four answer choices.

Each general knowledge question that was included in any part of the experiment was pre-tested in the same subject pool to ensure that the composition of questions in terms of difficulty and the field of general knowledge tested was comparable across parts of the study.[21]

At the beginning of the second stage of Part 1, groups were introduced. The computer randomly selected three participants from the same session to form a group. These groups remained fixed for the rest of the experiment. When groups were introduced, each group member saw the portraits of everyone in his group.[22] The pictures were on display for 30 seconds when no instructions were read and there was no option to advance to the next screen. In this way, the timing of when participants saw their other group members for the first time was held constant across all experimental conditions.[23]

In this second stage of Part 1, participants had to answer as many general knowledge questions as possible under some time pressure. Correct answers were paid the same fixed piece-rate as before (1.5 CHF).

Participants had 3 minutes to answer a series of 20 multiple choice questions. Questions appeared on their computer screens one at a time and an answer had to be submitted for the next question to appear on the screen. All participants saw the same sequence of the same 20 questions.

When the three minutes had elapsed, group members were automatically advanced to a feedback screen, whether or not they had provided an answer to all questions. This is the point at which the treatment manipulation occurred, as subjects received different information about their relative performance.

4.3. Experimental Conditions

When they performed the timed general knowledge test, group members knew what type of feedback they would receive. The feedback screen was on display for one minute and participants were unable to manually advance to the next screen.

In the baseline condition, each group member found out how many out of the 20 questions he answered correctly. No information on the performance of others in the group was provided. Thus, group members had no reference point against which to compare their general knowledge score.

In the private rank feedback treatment, a group member also discovered how his performance compared to the performance of others in his reference group. He found out whether

[21] The objective of the pre-tests was to include questions in the main experiment that would be neither too difficult nor too easy, that questions would be comparable across parts and that there would be no gender differences in performance on average. Average performance data across the different general knowledge tests, i.e., across Part 1 Stages 1 and 2 and Part 2, and by gender show that all of these objectives were fulfilled.

[22] Pictures were taken at the beginning of the session by an experimenter.

[23] This is important since Buser and Dreber (2016) find suggestive evidence that a simple group prime may activate norms of competition. In keeping this aspect constant, I rule out that the public rank feedback treatment, which shows subjects pictures of their other two group members, operates through priming the group more strongly compared to the other two conditions.

he ranked first, second or third in his group. Rankings were based on the number of questions answered correctly during the timed task with ties broken at random. This treatment introduced a performance competition and manipulated self-image in knowledge, relative to group members.

In the public rank feedback treatment, the feedback screen displayed the picture, the participant number and the performance rank of each group member. This way, the relative performance of each individual was common knowledge among the three members of a reference group. Therefore, the public rank feedback treatment made social image in intelligence explicit, while keeping the information about own relative performance the same as in the private rank feedback condition. With this condition, I can assess whether the perceived competition in reference groups is stronger when social image in intelligence is made explicit.

Table 2 summarizes the information shown on the feedback screen in each condition and Figures A1-A3 in the Appendix reproduce images of the feedback screens as they were shown to participants.

A control treatment introduced relative pay in an environment that was otherwise identical to the private rank feedback condition. The best performing group member on the timed general knowledge test received a substantial bonus of 5 CHF (≈5 USD) in addition to the piece-rate that was paid for each correct answer. Thus, the bonus amounted to 25% of the maximum earnings that a group member could receive for this timed general knowledge test. At the feedback stage, a participant observed his own performance rank and whether or not he would receive an additional bonus payment (see Figure A4). With this condition, I can test to what extent, *ceteris paribus,* any results change with the domain of relative concerns, that is, when money is or is not involved.

Table 2. Experimental Conditions

Baseline	Absolute performance feedback after timed general knowledge test
Private rank feedback treatment	Baseline + private information about own performance rank in group
Public rank feedback treatment	Baseline + public information about everyone's performance rank in group
Relative pay treatment (control)	Private rank feedback + best performer on timed test earns an additional bonus of 5 CHF

4.4. Part 2—Measuring Group Morale

The feedback that subjects saw at the end of Part 1 was also summarized in their Part 2 decision screens.

In Part 2, group morale was measured in a task that was independent of the relative performance competition in Part 1. Part 2 comprised 10 new multiple-choice general knowledge questions. Each group member had to provide an answer to each question. The piece-rate for correct answers stayed the same as in Part 1, that is, 25 points or 1.5 CHF.

For each question in Part 2, a participant had the option to share his answer with the other two group members. I chose this type of task explicitly to model the kind of helping behavior that takes place in workplace settings, where someone who knows information (how to accomplish a task, the needs of a particular client) can share this information with others to help their productivity.

Whenever a participant shared a correct answer to a question, the computer automatically replaced the incorrect answer of each group member who did not answer that question right with the correct answer that was shared. Sharing incorrect answers had no positive or negative effect on others in the group.

Accordingly, the total benefit to sharing an answer was either 0, 25 or 50 points (0, 1.5 or 3 USD)—depending on whether a correct answer was shared and on how many group members did not get a question right. This benefit went to *others* in the group.

Sharing an answer to a question cost a participant 1 point or 0.06 CHF (≈0.06 USD). Thus, when a group member shared an answer he was willing to invest 1 point in the performance of others.

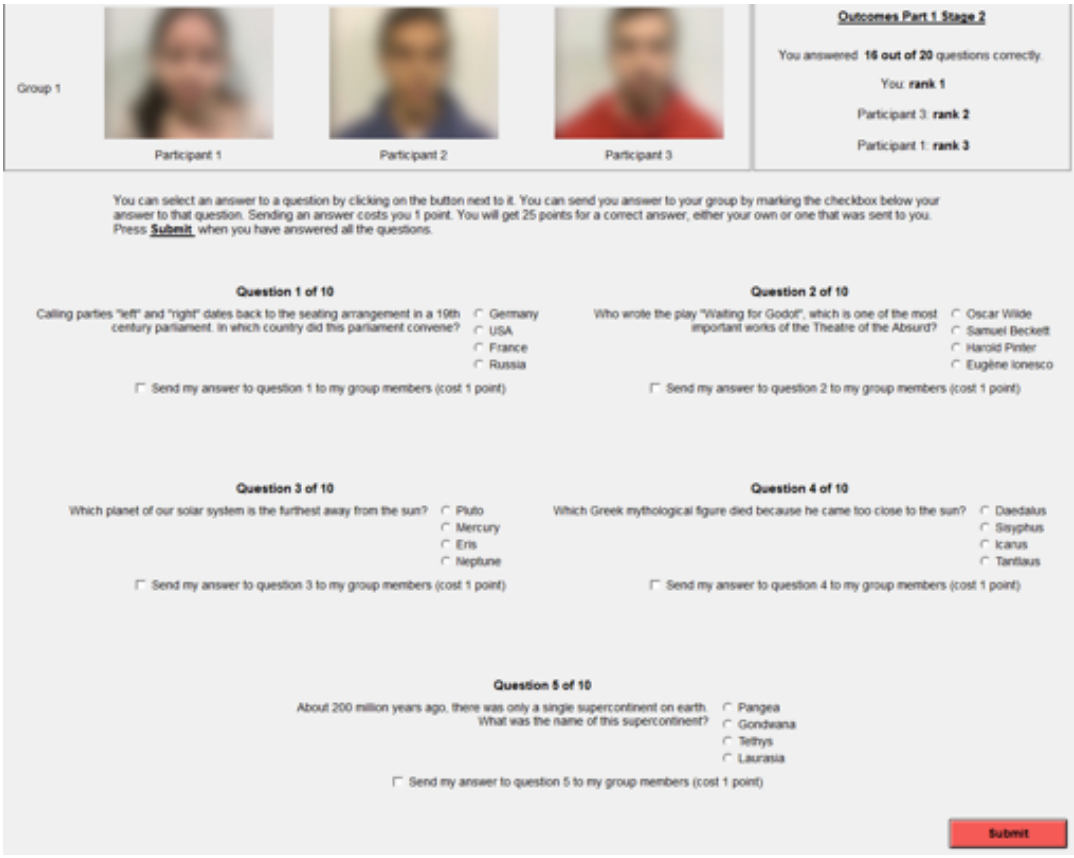
This way, I elicited participants' willingness to help others to perform better, observing 10 decisions for each person regarding whether or not to share the answer to a Part 2 question with others in the group. Figure 1 shows a screenshot of the decision screen. All instructions for participants described the act of sharing answers with the more neutral term "sending answers" to other group members.

At the very end of the experiment, group members found out how many Part-2-questions they answered correctly when they saw their summary of earnings in the experiment. This is the only feedback they obtained about choices made in Part 2. Participants were aware of this when they made their help decisions. Importantly, participants did not see who helped them, how much help they received or how many of their answers were replaced. I gave no feedback on help decisions to rule out that group members could seek to enhance their social image in a second dimension, namely, a reputation for being pro-social. This could interact with the rank treatments making it more difficult to isolate the direct effect of relative performance evaluations on the willingness to invest in the productivity of others.

When participants performed in Part 1 they did not know anything specific about the other parts of the experiment. Importantly, participants were unaware that Part 2 would entail pro-social choices. This allows me to rule out that any effect of rank feedback on the willingness to help other group members is driven by selection into high rank positions based on social

preferences, similarly to what Erkal, Gangadharan and Nikiforakis (2011) find in a study on competing for money and the subsequent willingness to redistribute earnings within a group.

Figure 1. Help Decision Screen (Public Rank Feedback Condition)



Notes. In all conditions, this screen displayed the portrait of every group member. Across conditions, the help decision screens only varied in the summary of performance on the timed task of Part 1, displayed in the box in the upper right corner of the screen.The screen of the *baseline* condition did only show how many question a participant answered correctly. In the *private rank feedback* condition the box also showed the performance rank of the participant who was looking at that screen. In addition to this information, in *the relative pay* condition, the participant who was looking at the screen also found out whether or not he had obtained an additional bonus payment.

4.5. Part 3—Beliefs

In Part 3 of the experiment, I elicited three beliefs for each Part-2-question; the subjective probabilities that a participant assigned to his answer and to the answer of each of the other two group members being correct.

I implemented a mechanism to elicit subjective probabilities in an incentive compatible way that was described in Karni (2009) closely following the experimental protocol introduced in Coffman (2014). In this part of the experiment, there were 100 lotteries available that had an integer-probability on [1,100] of selecting a correct answer to a question. In other words, there were lotteries that had a 1% chance, 2% chance, 3% chance ... up to a 100% chance to provide

a correct answer to a question. For each question, one of these lotteries was randomly selected, with each lottery equally likely to be chosen. Participants selected a threshold, X, such that for any lottery that selects the correct answer with a probability X or lower they would prefer their own answer to be evaluated for payment and for all lotteries that select the correct answer with a probability X or higher they would prefer the lottery to answer for them. Therefore, given a cut-off probability X', a participant believes that the answer he provided to a question is correct with probability X'.

For each question, participants stated three different such cut-off probabilities: one for the answer they provided themselves and one for the answer provided by each of the other group members. For payment, one of each of the three "types" of belief (self, two other group members) was randomly selected and evaluated. In this part, participants earned 2 CHF (≈2 USD) if they submitted a correct answer, regardless of whether this answer was submitted by them, by one of their group members or by a lottery.

For this belief elicitation task, the order of Part-2 questions was randomized at the subject-level.

With this data, I can assess whether participants intended to use help decisions instrumentally, to aid group members. Sharing answers that one believes to not know is not helpful, neither is it to share answers that one believes the others to know for sure. Moreover, I can use this data to control for the pure information effect of rank feedback through which the treatment may systematically affect the willingness to help compared to the baseline with no relative performance information.

I also asked participants to state how much help they expected to have received from each of the other two participants in their group. Participants earned 1 CHF when their guess of the number of answers that a group member shared with the group was within a margin of +/- 1 question to the actual number of questions this person sent. Expectations about others' help decisions could also influence the willingness to share knowledge, out of a desire to reciprocate or a desire to conform. Since rank feedback could also influence the help a group member expects to receive from others, it is important to have data on expected help.

4.6. Part 4—Questionnaire

In an exit survey, I collected several measures to assess perceptions of social relations in the experimental reference groups. I measured the intensity of perceived competition in the experimental (reference) groups with an agreement to the following statement: "I felt in competition with the other two members in my group when performing this task." on a 9-point Likert scale. "This task" refers to the timed general knowledge test. This data was collected about 30 minutes after this stage of the experiment and serves as a test of whether relative

performance feedback produced a sense of competition among the members of the reference group.

On the same scale, participants also answered questions to assess to what extent performing well on the general knowledge test and being seen performing well by others was desirable and to evaluate whether they thought that the questions actually tested general knowledge.

In addition, I measured the social distance among members of experimental reference groups with the Oneness index. It records the subjective perception of closeness between a participant and each of his group members, ranging from no connection at all to feeling "at one" with another person. This scale is widely used in psychology to measure the closeness of social relationships (Gächter, Starmer, and Tufano 2015a) and predicts behavior in economics studies involving decision-making in groups (see, e.g., Gächter, Starmer, and Tufano 2015b). The Oneness index is constructed from responses to the Inclusion of the Self in Other (IOS)-scale and the we-scale. On the IOS-scale, a participant indicated how close he felt to another group member by selecting a pair of circles that best represents the relationship with that other group member. In these pairs, one circle depicts the participant and the other circle the other group member. Across the pairs, the circles differ in how much they overlap. See Figure A5 for the pictogram used in this task. For the we-scale, a participant indicated on a 7-point Likert scale to what extent he would use the term "we" to characterize himself and another group member. The Oneness index is simply the average of a participant's responses on these two scales for a given group member. As a measure of social relations in reference groups, this outcome variable provides complementary evidence for the change in competitiveness from the baseline to the treatment conditions.

To complement my behavioral measure of group morale, I also elicited general attitudes toward cooperation, toward working in groups or working alone and toward competition, following the procedure for eliciting general attitudes presented in Duffy and Kornienko (2010). I also took their set of items eliciting attitudes on competition. For each of these general attitudes, participants evaluated the extent to which four statements applied to them on a 9-point Likert scale ranging from 1 -does not apply at all, to 9 -definitely applies.[24] For each category, e.g. for cooperativeness, an index is constructed which is the average response to the four statements that belong to the category, reverse-scoring responses when necessary. A list of all four items for each category with summary statistics is in the Appendix (Table A2).

Lastly, I elicited positive and negative reciprocity as general traits with the set of questions described in Falk et al. (2016). The authors provide evidence that responses to these questions are highly correlated with behavior in experimental games that are typically used to measure

[24] For example, participants rated "I like to share my ideas and material with others." which is a statement on cooperation or "I find that working in groups is often inefficient" which is an attitude toward working in groups. I included four questions per attitude to reduce the influence of an idiosyncratic question.

reciprocity with decisions involving real monetary stakes. This allows me to test whether the treatment—that may impact pro-social help—also extends to general pro-social attitudes.

The questionnaire concluded with a few questions on demographics and an elicitation of attitudes towards risk and towards competition. For each of these attitude measures, participants were asked to position themselves on a scale from 0 (very risk-averse; not competitive at all) to 10 (very risk-seeking; very competitive).

4.7. Hypotheses About Behavior in the Experiment

The experiment is designed to assess whether relative performance information lower group morale when participants make 10 help decisions in Part 2. I hypothesize that relative performance evaluations put group members in a competitive mindset with respect to one another, compared to an environment that, *ceteris paribus*, does not provide this information:

> Hypothesis 1: Rank feedback causes a sense of competition in the reference group in Part 1.

I test the mechanism that this competition activates positional concerns over relative performances, which may lower the motivation to subsequently extend pro-social help to others in order to increase their productivity (see also section 3). My study offers a direct test of the null hypothesis,

> Hypothesis 2-0: The sense of competition under relative performance feedback does not lower the willingness to help other group members perform better in Part 2.

against the alternative hypothesis that,

> Hypothesis 2-A: The sense of competition under relative performance feedback lowers the willingness to help other group members perform better in Part 2.

4.8. Experimental Procedures

The experiment was conducted in English at the Laboratory for Experimental and Behavioral Economics at the University of Zurich. In total, 282 participants, most of them students at the University of Zurich and the Swiss Federal Institute of Technology in Zurich, took part in the experiment. Table 3 lists the number of participants in the different conditions.[25]

[25] While 7 out of 9 sessions comprised 24 participants (8 groups), two sessions in the public rank feedback condition were conducted with 21 participants (7 groups) because some of those who registered did not show up.

Table 3. Overview of Data

Condition	Participants
Baseline	obs.=72, 24 groups
Private rank feedback	obs.=72, 24 groups
Public rank feedback	obs.=66, 22 groups
Relative Pay	obs.=72, 24 groups
Total	obs.=282, 94 groups

At the beginning of a session, an experimenter took pictures of all participants before participants took their seat in the laboratory. Participants were called individually by their participant number and were instructed to make a neutral face for the portrait. The composition was the same for every portrait, with a zoom on the face leaving out the upper body. Participants gave informed consent to having their picture taken and to the fact that these pictures may be linked to some of their choices in the experiment.

The instructions for the study were displayed on the computer screen in a participant's cubicle. Screenshots of the instructions and decision screens exactly as they were shown to participants are reproduced in the Online Appendix. An experimenter read the instructions for a part out loud just before participants made choices in that part. Before Part 2 and Part 3, participants also answered comprehension questions and the study only advanced after all participants had answered the questions correctly.

The experiment was programmed in z-Tree (Fischbacher 2007). The computer selected for each participant whether the first stage of Part 1 or the second stage of Part 1 was selected for payment, giving equal weight to each option. Earnings in Parts 2 and 3 were always paid out. Average earnings were 40.00 CHF (≈40 USD) (including a 15 CHF show-up fee).

5. Results

First, I will consider results in support of Hypothesis 1. Then, I will turn to results on the treatment effect of relative performance information on group morale that lead to the main result of the paper regarding Hypotheses 2-0 and 2-A. I will then briefly present findings on the factors that predict helping behavior and consider findings from a control treatment that introduces relative pay. I conclude this section with results on the link between group morale and group productivity in the environment under study.

5.1. Do Relative Performance Evaluations Affect Perceptions of Competition?

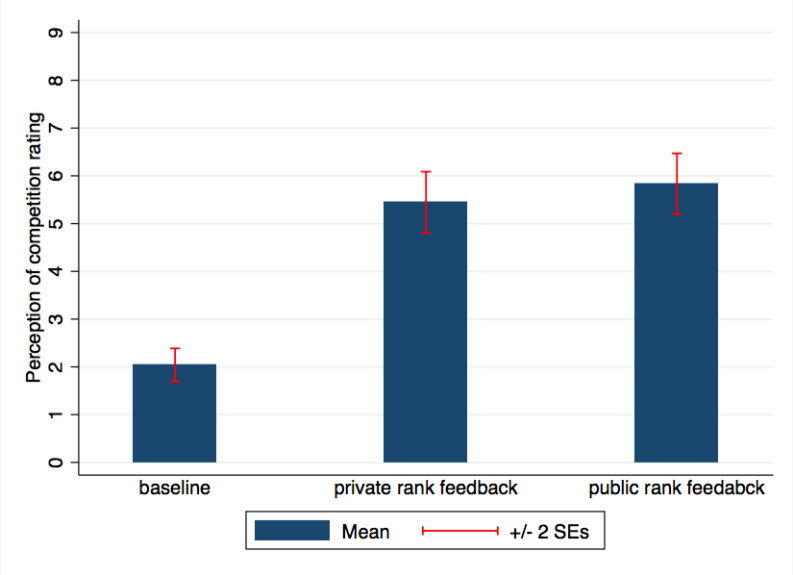
Rank feedback on the timed general knowledge test of Part 1 mirrored actual performance differences among the group members in the absolute majority of reference groups. Performance on the timed general knowledge test of Part 1 varied substantially and performance ties occurred only in 12 % of the randomly formed experimental groups. On average, participants answered 11.3 out of 20 questions correctly (SD=3.0). The best performers in my sample answered 17 questions correctly and the worst performers 3. The empirical distributions of performance on this timed task are very similar across the four experimental conditions (see Figure A6 in the Appendix) and a Kruskal-Wallis test fails to reject the null hypothesis that these performance samples are drawn from the same population (p=0.4346).[26] This is expected given that the task tested existing knowledge and effort had little to no scope to increase performance on this test.

Hypothesis 1 states that relative evaluations on this first timed general knowledge test produces perceptions of competition between group members. To address this question, I turn to data from the post-treatment questionnaire, in which several items asked specifically about the timed task in Part 1. I focus on the causal effect of providing relative performance feedback on participants' agreement with the statement "I felt in competition with the other two members in my group when performing the task." Responses can range from 1 (does not apply at all) to 9 (definitely applies). Figure 2 visualizes the marked differences in the perception of competition in groups between the baseline and the two rank feedback conditions. In the baseline condition, the average sense of competition in groups is 2.0 on this competitiveness scale which is very low. The estimated average treatment effect of private rank feedback is an increase of 3.4 points on the 9-point competitiveness scale with a 95% confidence interval of [2.68, 4.12]. The estimated effect of public rank feedback is a 3.8-point increase in the sense of competition with others in the reference group with a 95% confidence interval of [3.10, 4.49]. These effects are highly statistically significant (two-sided t-tests: Bas-Priv. RF $p<0.0001$; Bas-Pub. RF $p< 0.0001$). The fact that this data was collected 30 minutes after participants completed Part 1 indicates that this information had a lasting impact on the sense of competition in reference groups.

Whether the performance feedback is public or private appears to not make a sizable difference in the sense of competition in reference groups. The effect of a public ranking over a private ranking is 0.40 points on the competitiveness scale with a 95% confidence interval of [-0.51, 1.29], which includes 0 and is not statistically significant (two-sided t-test p=0.3926).

[26] Consistent with this, participants did also not perform systematically differently across the four experimental conditions on the first set of questions that was administered before the treatment manipulation occurred: Kruksal-Wallis test p-value=0.5410. On average, participants provided correct answers to 61% of questions on this first untimed general knowledge test (mean number of questions correct=6.160, SD=1.640).

Figure 2. Perceptions of Competition by Condition



Importantly, while relative performance feedback changed the perception of competition among group members, other aspects of how individuals evaluated the task in Part 1 did not change. Using a 9-point scale on which higher numbers indicate more agreement with a statement, participants thought that the general knowledge questions did, in fact, measure their general knowledge (mean agreement=7.26, SD=1.83). Participants also wanted to perform well at the Part 1 performance stage (mean agreement=8.15, SD=1.32) and would be impressed if others answered 90% or more of the general knowledge questions in the experiment correctly (mean agreement=7.35, SD=2.12). The treatments had virtually no effect on this (see Figure A7 in the Appendix for a visualization). All in all, this supports the interpretation that the timed task of Part 1 is relevant for perceptions of competence, and that individuals valued performing well and being seen as performing well on the task.

Relative performance evaluations also increase the perceived social distance between members of experimental reference groups. The Oneness index (Gächter, Starmer, and Tufano 2015b) is a simple instrument to measure how close subjects perceive themselves to be to every other group member on a scale from 1 (no connection at all) to 7 (feeling as "one" with another person). In column 1 of Table 4, I present the results from ordinary least square (OLS) regressions of the Oneness index, averaged over the two group members of a participant, on treatment indicators. The results indicate that rank feedback decreased the social closeness between members of experimental reference groups, when group members were asked 35 minutes after the end of Part 1 how close they felt to others in their group. The estimated average treatment effect is 0.51, or half a category on this 7-point scale, under private rank information and 0.40 under public rank information.

In contrast, *general* attitudes toward working in groups or working alone were not systematically affected by the treatments (columns 2) and 3) of Table 4).

Table 4. Rank Feedback, Relative Pay, Social Closeness and General Attitudes on Groupwork

	Social closeness with group members	General attitude working in groups	General attitude working alone
Private RF	-0.514***	-0.243	0.191
	(0.181)	(0.229)	(0.187)
Public RF	-0.399*	0.155	0.071
	(0.206)	(0.215)	(0.200)
Relative Pay	-0.469**	-0.215	0.184
	(0.205)	(0.234)	(0.230)
Constant	3.028***	5.083***	6.038***
	(0.130)	(0.154)	(0.122)
Obs.	282	282	282

Notes. The data was collected for each participant i in a questionnaire at the end of the study. *Private RF* and *Public RF* are indicators for participant i privately observing his performance rank on the timed task or publicly observing the performance rank of everyone in his group, respectively. *Relative Pay* indicates that participant i was in the control treatment that paid a bonus to the best performer, in addition to providing rank feedback in private. Social closeness with group members is the average of the two responses of i on the oneness index measuring how close i feels to each group member j. This variable ranges from [0,7]. General attitudes on working in groups and on working alone are indices that range from 1 (strongly negative attitude) to 9 (strongly positive attitude). Robust standard errors are in parentheses, 70 group clusters allow for correlated observations at the group and at the subject level. *Significant at the 10% level,** at the 5% level, *** at the 1% level.

All these findings support the conclusion that relative performance evaluations changed the perceptions of social relations in experimental reference groups. This leads to the following first result:

> Result 1: Rank feedback information causes perceptions of competition between group members and increases their social distance.

There is strong evidence that the relative performance information treatments successfully manipulated the sense of competition, thereby increasing the salience of relative performance concerns in reference groups. I do not find evidence that publicly ranking group members further increases relative performance concerns in reference group, compared to a scenario in which group members observe their performance rank privately. How does the increased salience of relative performance concerns under relative performance evaluations affect group morale, that is, the willingness of group members to help each other perform better? These are the next results I turn to.

5.2. Do Increased Perceptions of Competition Affect Group Morale?

In the following section, I present results from analyzing help behavior in Part 2 of the experiment with a focus on the treatment effect of relative performance feedback. Throughout this section, I compare behavior in the baseline condition to behavior in the two treatments that provided relative performance information, that is, the private rank feedback and public rank feedback treatment conditions.

First, I estimate the causal effect of relative performance evaluations on the total number of answers a group member shared out of the 10 Part 2 questions (see Table 5).

Table 5. Average Treatment Effect of Rank Feedback on Help

ATE $\boldsymbol{\mu_{bas} - \mu_{treat}}$	95% confidence interval
Bas-Priv. RF: 0.36	[-0.69, 1.41]
Bas-Pub. RF: 0.013	[-1.12,1.15]

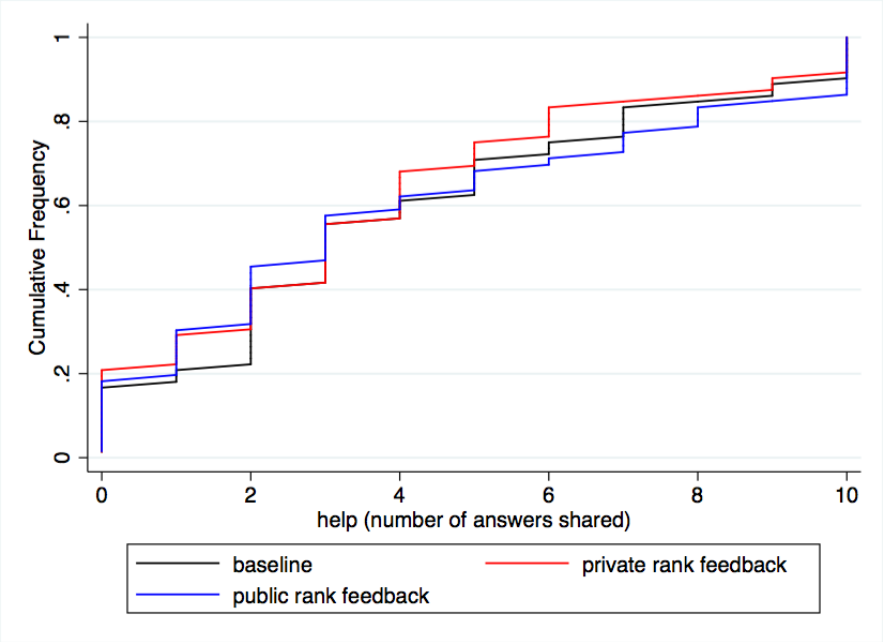
The estimated average treatment effect of private rank feedback is a decrease in 0.36 questions shared, with a 95% confidence interval of [-0.69,1.41] and a p-value of 0.50. This confidence interval includes zero and the decrease in average help of 3.6% is not statistically significant at any reasonable significance threshold. The estimated average treatment effect of relative performance feedback is even smaller, a decrease in 0.013 questions shared with a confidence interval of [-1.12, 1.15] that includes zero and a p-value of 0.98.[27] This is not sufficiently strong evidence against hypothesis H2-0, that relative performance evaluations have no effect on help behavior in reference groups. It is important to highlight that the estimated confidence intervals are narrow, which means that the average treatment effects are precisely estimated. This suggests that the failure to reject hypothesis H2-0 is not driven by measuring group morale with substantial noise. Section 6 gives a detailed account of the statistical power of this study.

Average help is an important summary statistic, in particular in the face of the sharp directional hypothesis H2-A on how relative performance concerns affect group morale in reference groups. Figure 3 shows that also the empirical distributions of number of answers shared look very similar across the three experimental conditions. A Wilcoxon-Mann-Whitney test fails to reject the null hypothesis that the baseline and treatment samples of help behavior are drawn from the same population (Bas-Priv. RF p=0.535, Bas-Pub. RF p=0.775).

[27] These are the Average Treatment Effects (ATE) and confidence intervals when I treat each group member as one observation. The precision of my ATE estimates are very similar when I collapse observations at the group level, which takes into account that group members may have been exposed to some common shocks. The 95%-confidence intervals are Bas-Priv. RF [-0.65,1.37] with a p-value of 0.477 and Bas-Pub. RF [-1.04,1.07] with a p-value of 0.981.

The empirical distributions of help also visualize that there are three different empirical "types" of people. Some people are, what I call, "resolute helpers" who share all ten answers to their questions. There is also the "selfish"-type who never shares her knowledge with others in the group. The majority of people share some answers. In the baseline condition, for example, 11% are resolute helpers and 17% of group members are selfish. This is reassuring since to be able to identify a negative effect of relative performance concerns on group morale, it is important that the baseline share of the selfish-type is not too large. The shares of the selfish-type and the resolute helpers are remarkably consistent across the three experimental conditions (see Figure 3), which is further suggestive evidence that the treatments did not systematically affect behavioral motives for knowledge-sharing.

Figure 3. Empirical CDF of Help by Condition



So far, there is no strong evidence in support of hypothesis H2-A, that relative performance concerns, activated by relative performance evaluations, have a sizable negative effect on group morale.

However, with this type of statistical hypothesis testing one cannot control for the fact that beliefs about the value of own help to the group may be affected by relative performance information, which may counteract a negative treatment effect.

I find no evidence that self-confidence in answering general knowledge questions correctly is affected by rank feedback information.[28] But I find that knowing performance ranks has a small but measurable impact on the confidence in the ability of the other two group members

[28] Table B1 in the Online Appendix reports the results from OLS regressions that predict, conditional on performance rank, the percent chance that a participant assigns to himself having provided a correct answer to a Part 2 question as a function of treatment indicators and controls.

to answer general knowledge questions correctly in the expected direction.[29] Taken together, these results imply that, for example, rank 1 group members who know their rank deem their help to be slightly more valuable, on average, than rank 1 group members in the baseline condition. This can potentially counteract a negative effect of relative performance concerns.

Therefore, I control for these beliefs in binary choice models that predict the willingness to share answers. I estimated the following general model:

$$Prob(share_{ki}) = g\big(\beta_0 + \beta_1 I_{privRF} + \beta_2 I_{pubRF} + \beta_3 belief\ self\ correct + \beta_4 belief\ others\ in\ group\ correct + [Controls_i + FE_k]\big)$$

$share_{ki}$ indicates whether participant i shared his answer to question k with the others. I_{privRF}, I_{pubRF} indicate whether this participant privately observed his performance rank on the timed task or publicly observed the performance rank of everyone in his group. The variable *belief correct self* is the subjective probability that group member i gave to his answer to question k being correct. The variable *belief others in group correct* is the average of the probabilities that group member i gave to his two group members having provided a correct answer to question k. I introduce the other control variables below when I briefly consider results on the factors that correlate with helping decisions (section 5.3). I report results from Probit and Linear Probability Models.

Table 6 columns 1)-3) show the results from fitting a Probit model. The first specification predicts the willingness to help as a function of treatment indicators with no further control variables. The predicted marginal effect of private rank feedback or public rank feedback on the willingness to help are very small and statistically insignificant (private RF: -0.036, public RF: -0.001) which simply confirms the previous findings.

The second specification adds controls for beliefs about correct answers. The willingness to share the answer to a question increases in the probability that a group member assigns to his answer being correct (*belief correct (self),* predicted marginal effect at means of covariates=66 percentage points) and decreases as he gives a higher probability to his group members having provided a correct answer (*belief correct (others),* predicted marginal effect at means of covariates=-28 percentage points). Both effects are highly statistically significant. The signs of the two coefficients, positive for *belief correct (self)* and negative for *belief correct (others)*, indicate that participants help more when they believe their help will be valuable. [30]

[29] Table A1 in the Appendix presents results from OLS regressions that predict the percent chance that a participant assigns to another group member providing a correct answer to a Part 2 question, conditional on the rank of the participant who makes this judgement, treatment indicators and further control variables. The results indicate that relative performance information systematically affects the confidence in other group members' likelihood of providing correct answers in the expected direction, for rank 1, rank 2 and rank 3 group members in the private and the public rank feedback conditions. For example, observing their relative performance rank in private (public) reduces rank 1 group members' confidence in their other group members' ability to provide correct answers to Part 2 questions by about 5.5 (7.2) percentage points.

[30] In the Online Appendix, I also present results from models in which I relate these belief variables, instead of testing for them separately (Table B3). I construct a control variable *valuable help* that ranges from [0,2]. It sums the two conditional probabilities that a participant i assigns to her answer improving the performance of group member 1 and

This evidence is consistent with the interpretation that participants share answers in order to improve the performance of others in their reference group.

Importantly, however, the introduction of these control variables does not change the inference about the treatment effect of relative performance evaluations on the willingness to help others perform better, the predicted marginal effects remain very close to zero (private RF -0.038, public RF 0.006).

Table B2 in the Online Appendix presents results from specifications similar to the ones presented in columns 1-3 of Table 6 but pooling the data from the private and public rank feedback treatments. This further increases the power to detect a treatment effect. I pool the data from these two treatments for additional robustness analysis and because they had very comparable effects on the perceptions of competition in Part 1 (see Figure 2). These results also fail to reject hypothesis H2-0 that the performance competition of Part 1 has no effect on the willingness to help in Part 2.

I also perform additional robustness analysis in which I add terms that interact each treatment indicator with the variables *belief correct (self)* and *belief correct (others)* in a linear probability model. This directly tests whether participants are less likely to share the answers that they think are going to be valuable to others under either type of relative performance feedback. While the estimated sign of the interaction terms is typically negative—consistent with hypothesis H2-A—they are far from reaching statistical significance (see Table B4 in the Online Appendix).

Heterogeneity in Treatment Effect by Rank in Competition

Looking at average treatment effects in the whole sample could mask substantial heterogeneity in how strongly group members of different performance ranks change their help behavior after rank feedback. Figure 4 displays average helping by rank and by condition. For rank 1 and rank 2 group members, average help is slightly lower under relative performance evaluations, whereas it is slightly higher for rank 3 group members.

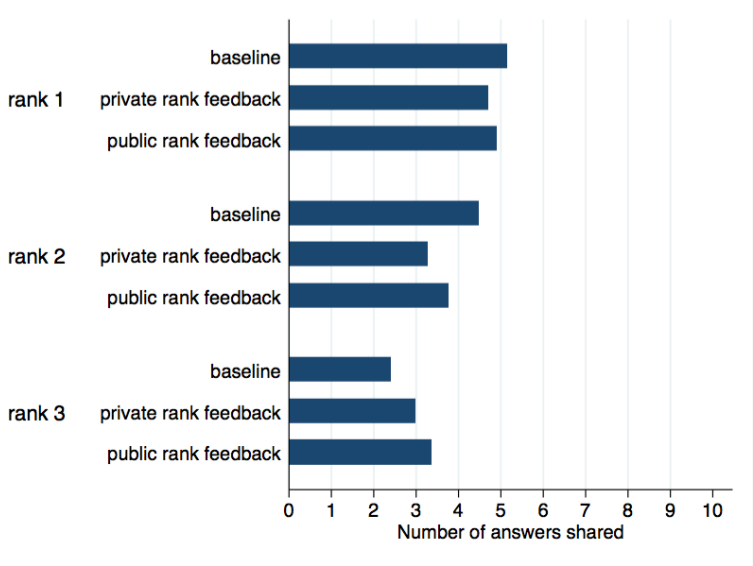
group member 2, conditional on the group member not knowing the answer to that question. I can back out these probabilities from the three beliefs about correct answers that participant i stated for each Part 2 question under the assumption that they are independent. None of the results presented in the main text change when I control for the expected value of own help to others in this way.

Table 6. Predicting the Willingness to Help

	Predicting Prob(share answer to question k)			
		Probit		OLS
Private RF	-0.03623	-0.0388	-0.0132	-0.00560
	(0.0500)	(0.0562)	(0.0485)	(0.0330)
Public RF	-0.0012	0.006	0.0366	0.0230
	(0.0522)	(0.0571)	(0.0458)	(0.0310)
Belief correct (self)		0.659****	0.7803****	0.558****
		(0.0947)	(0.1314)	(0.0654)
Belief correct (others)		-0.2762**	-0.3682***	-0.269****
		(0.1126)	(0.1404)	(0.0767)
Actual correct			-0.0088	-0.00856
			(0.0342)	(0.0225)
Performance part 1			0.1563	0.125
			(0.1506)	(0.0998)
Expected help			1.4436****	1.053****
			(0.1776)	(0.0661)
Risk attitude			-0.2478****	-0.150***
			(0.0768)	(0.0552)
Female			-0.0294	-0.0234
			(0.0487)	(0.0340)
Question FE	No	No	Yes	Yes
Constant	Yes	Yes	Yes	-0.334****
				(0.0892)
Obs.	2100	2100	2100	2100
(pseudo) R^2	0.0009	0.08	0.35	0.381

Notes. Predicted marginal effects at mean level of covariates are reported (Probit). *Share answer to question k* is an indicator for whether participant shared the answer to a Part 2 question k with others. *Private RF* and *Public RF* are indicators for participant i privately observing his performance rank on the timed task or publicly observing the performance rank of everyone in the group. *Belief correct self* ranges from [0,1] and is the subjective probability that participant i gives to the event that his answer to question k is correct. *Belief correct others* ranges from [0,1] and is the subjective probability that participant i gives to the event that his average group member provided the correct answer to question k. Actual correct indicates whether participant i provided the correct answer to question k. *Performance Part 1* records the fraction of questions that participant i answered correctly during the timed task of Part 1. *Expected help* is the fraction of answers hat participant i thinks the two group members shared, on average. *Risk attitude* is where participant i positioned himself on a scale from 0=very risk-averse to 10=very risk seeking, divided by 10. Robust standard errors are in parentheses, 70 group clusters allow for correlated observations at the group and at the subject level. *Significant at the 10% level,** at the 5% level, *** at the 1% level, **** at the 0.1% level.

Figure 4. Help by Treatment and Rank in Competition (Means)



However, Wilcoxon-Mann-Whitney tests fail to reject the null hypothesis that there are any differences in how help, conditional on rank, is distributed comparing help behavior in the baseline condition with help behavior under either type of relative performance feedback (see Table 7 for a summary of p-values).[31] Within each condition, higher ranked group members typically provide more help, on average, than lower ranked group members, consistent with an interpretation that participants use the help technology because they seek to improve the performance of others in their group.

Table 7. Wilcoxon-Mann-Whitney Tests

rank	compare number of answers shared across conditions	p-value
1	Baseline-Private RF	0.6474
	Baseline-Public RF	0.7989
2	Baseline-Private RF	0.2118
	Baseline-Public RF	0.2973
3	Baseline-Private RF	0.6070
	Baseline-Public RF	0.728

[31] In the Online Appendix I report results from binary choice models that predict the willingness to help conditional on performance rank, treatment indicators and controls for beliefs about correct answers. For this analysis, the data from the private rank feedback and public rank feedback treatments are pooled (Table B5). Even when I pool the data, for this type of analysis, the sample size is relatively small. This makes the inference that one can draw based on this sample limited, though the relatively small magnitudes suggest modest effects of the treatment on the willingness to help also when I allow for heterogeneous treatment effects. The point estimates of the treatment effect tend to be small in magnitude, e.g. the linear probability model predicts a marginal effect on the willingness to help by 6-11 percentage points, and are never statistically significant.

Finally, there is another way to consider heterogeneity in treatment effects based on performance on the timed task of Part 1. Note that relative performance information could be particularly relevant for individuals who perform neither exceptionally well nor very poorly on the timed general knowledge test. The absolute performance feedback in the baseline condition may have provided participants whose scores were on the tails of the performance distribution, e.g. 4 questions or 16 questions answered correctly, already with a clear sense of how these outcomes compare to the performance of others, thereby activating relative performance concerns also in the baseline. In additional robustness analysis, I account for the fact that the value of the treatment information may differ according to one's absolute performance on the timed general knowledge task. Table B6 in the Online Appendix presents results of a Probit estimation that predicts the probability to share answers for subsamples of participants whose performance on the timed general knowledge test was neither exceptionally good nor bad according to performance percentiles.

This additional robustness analysis confirms the finding that relative performance information *per se* appears to not lower the intrinsic motivation to help others perform better.

This leads to the main result of the paper:

> Result 2: Hypothesis 2-0, that the performance competition under relative performance feedback has no effect on the willingness to help other group members perform better, cannot be rejected.

Relative Performance Evaluations, Expected Help and Cooperative Behavior in General

Consistent with this main result, the two treatment manipulations do also not extend to have effects on neither the expected help behavior of others nor the general willingness to act cooperatively. Table A3 in the Appendix presents results from OLS regressions that predict the number of answers that a participant believes to have received from the two group members, on average, as a function of treatment status (column 1). The estimated association is negative, though small in magnitude and not statistically significant. General cooperativeness is defined as the desire to help others in general. It is measured with a cooperativeness index constructed from a participant's agreement with four statements, for example, "I love to help others" (see Table A2 in the Appendix for all statements). While negative in sign, the coefficient is far from statistically significant, suggesting at most a weak relationship (see column 2 of Table A3). Participants' general pro-social inclinations in hypothetical situations involving positive and negative reciprocity are also not systematically affected by treatment. While linear regressions suggest a negative association between treatment and these measures, they are not reliably statistically significant (columns 3) and 4) of Table A3).

To summarize, while the private and public rank feedback treatments reliably change perceptions of social relations in reference groups, that is, the sense of competition and social

distance between group members, they do not systematically affect actual help behavior, expectations about the help behavior by other group members, the desire to help others in general or broader pro-social inclinations.

5.3. What Predicts Variation in Help?

While the previous section focused on the treatment effect of relative performance feedback on the willingness to help—finding no evidence against a null effect—it is instructive to also consider the variables that help to explain variation in the willingness to help in binary choice models. This analysis provides insights into what motivates participants to share answers with others. Column 3 of Table 6 shows results of a Probit estimation that predicts the willingness to help as a function of treatment indicators, performance controls, beliefs about correct answers, the expected help by others, risk attitudes, gender and question fixed effects.

We have already seen that *beliefs about correct answers* are highly predictive of helping decisions, which is consistent with the interpretation that participants share answers that they think will help other group members' performance. This lends support to the interpretation that the task in Part 2 elicits group morale, that is, subjects share knowledge in order to improve the performance of others.

The variable *expected help* ranges from [0,1] and is the fraction of Part 2 answers that a participant expected to have received from the two group members, on average. It explains variation in the willingness to help. The point estimates of the coefficient suggest that the expected help by others has a sizeable effect on the willingness to help them perform better. [32] This is consistent with the interpretation that behavior in Part 2 is motivated by pro-social inclinations, i.e. positive reciprocity in the form of mutual positive expectations in a group.[33] This provides further evidence suggesting that the task in Part 2 operationalizes group morale, whereby group members are more willing to help others who they think will also help them.

The coefficient on the measure of *risk attitude* suggests that individuals who describe themselves as risk-seeking are less willing to share their answers with others than individuals who describe themselves as risk-averse. This may indicate that one motive behind sharing answers with other group members is to insure them against a lack of knowledge on a particular question.

Conditional on this full set of controls, there are no *gender* differences in the willingness to share answers with others. In this context, women do not act systematically more pro-socially than men.

[32] Note that the estimated and predicted marginal effect of the variable "Expected help" *at the mean level of other covariates* exceeds one. The predicted *average marginal effect* of the variable "Expected help" is, of course, smaller than one (0.96025 with a standard error of 0.0669).
[33] An alternative interpretation is more mechanical, whereby participants base expectations of others' help on their own behavior.

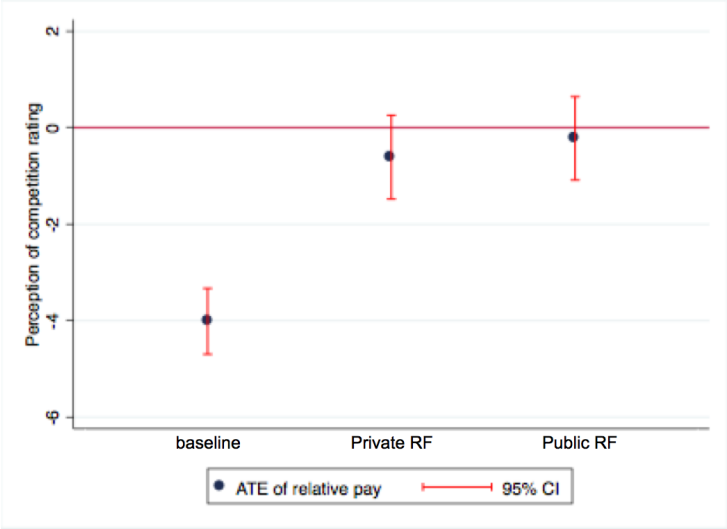
Lastly, we find no association between *absolute performance on the timed general knowledge test* and the willingness to help others in Part 2 of the experiment. In this decision context, the higher experimental earnings that group members were aware of when making help decisions did not translate into an increase in generosity.

The corresponding OLS estimates, that largely agree with the Probit estimates, are reported in column 4 of Table 6.

5.4. Relative Pay Treatment

Next, I turn to the results from a control treatment that entails relative pay.[34] With this data, I can directly compare the intensity of competition under relative performance information to the benchmark of a money competition.

Figure 5. Intensity of Competition Under Relative Pay Compared to the Three Other Experimental Conditions



Notes. This graph compares the average intensity of competition in reference groups in the relative pay treatment to reference groups in the other experimental conditions. It shows the estimated difference in mean agreement with the statement "I felt in competition with the other two members in my group when performing this task." on a scale from 1 ("does not apply at all") to 9 ("definitely applies") and its 95%-confidence intervals.

Consistent with earlier results, competing for relative pay has a large and significant increase (about 4-points on a 9-point scale) on the sense of competition in reference groups compared to the baseline in which relative concerns are absent. Moreover, I find no evidence

[34] This treatment is the same as the private relative performance feedback treatment in terms of relative performance information that group members obtain at the end of Part 1 of the experiment. The best performer of a group makes a substantial monetary bonus that amounts to 25% od potential Part 1 earnings.

that competing for relative pay increases the intensity of competition in reference groups, compared to otherwise identical environments in which group members only obtain relative performance information either in public or private (see Figure 5).

Results from an OLS regression indicate that relative pay also increases the social distance among members of a reference (see Table 4 column 1). Competing for relative pay reduces the reported social closeness to the other group members by -0.469 points on average. This is about half a category on the 7-point Oneness Index. Thus, competing for relative pay has a comparable negative effect, in terms of absolute effect size, on the perceived social closeness among members of a reference group (see section 5.1).

Taken together, these findings are consistent with the interpretation that the intensity of competition in reference groups is comparable across the domains of relative pay and relative performance comparisons.

Consistent with this first result, I also find no evidence that competing for relative pay has a sizable negative (or positive[35]) effect on the willingness to extend costly help to improve the performance of other group members. The average treatment effect of relative pay on help is 0.639 answers shared less after the experience of a competition for pay, with a 95%-confidence interval [-0.45, 1.72] that includes zero. Results from choice models that predict the willingness to help others perform better as a function of an indicator variable for relative pay and controls for beliefs about correct answers confirm this conclusion (Table A4 in the Appendix).

These findings provide further evidence that a competitive environment in and of itself does not systematically affect the willingness to extend costly help to improve the productivity of others.

5.5. Missed Opportunities to Help and Productivity

This paper sets out with the observation that it is important to apprehend what makes groups productive, in order to obtain a better understanding of the factors that determine productivity in firms. Next, I zoom in on the counterfactuals that are seldom observed outside of the laboratory. I identify the missed opportunities to help other group members perform better and quantify what they imply for group productivity and efficiency.

A missed opportunity is defined as an instance in which one group member submitted the correct answer to a question but did not share the knowledge, with the result that another group member ended up submitting the wrong answer to a question. Each group in my experiment had, on average, 4.0 of such missed opportunities (SD=2.7).[36] Group productivity would have

[35] In the relative pay treatment, the Part 1 earnings of the best performer are at least 25% higher compared to the other two group members. I find no evidence that this has any effect on the winner's generosity to extend costly help, compared to the best performers in the baseline condition (ranskum test p-value=0.9583).

[36] Group members who did not share their answers, although they got it right and sharing could have improved the performance of at least one group member, had an average confidence of 64% that their answer was correct; in 50% of these cases group members gave a chance of at least 70% to their answer being correct. This indicates that the

increased by 18%, on average, had these answers been shared. The average efficiency loss of missed opportunities—unrealized group earnings under higher productivity net the cost of help—is 5.8 CHF (≈5.8 USD) (SD=4.0). Table A5 also reports these numbers separately for each experimental condition.

This analysis shows that a change in group morale, a decrease or increase, would have real consequences on group productivity in the environment under study.

6. Discussion & Conclusions

The study investigates whether relative performance evaluations lower group morale in reference groups. In a laboratory experiment, I vary by treatment whether or not members of a reference group obtain relative performance feedback on a timed general knowledge test. Rank feedback is either given in private or in public. I then test how the salience of relative performance concerns under relative performance feedback spills over to the willingness to help others in the reference group perform better.

My data show that relative performance evaluations on a task that people perceive as relevant for their social and self-image cause a large and lasting increase in the sense of competition in reference groups, compared to the baseline condition in which no relative performance feedback is provided. The intensity of competition under private or public rank feedback is comparable to how strongly members of reference groups compete for relative pay in a control treatment. Yet, I find no evidence that relative performance concerns from the competition spill over to subsequent help behavior. Relative performance evaluations and relative pay do not systematically affect the willingness to help others in the group. Beliefs about the value of own help in terms of improving other group members' performance are a strong predictor of actual help behavior, which is consistent with the interpretation that participants share knowledge in order to help the performance of others.

These findings provide initial evidence suggesting that the intrinsic motivation to extend on-the-job help does not systematically decrease as the work culture becomes more competitive.

How credible is my main finding that I fail to reject hypothesis H2-0 of no treatment effect on help behavior in reference groups?

The answer to this question is linked to considerations of statistical power. Table 8 reports the power of this study to detect the effect that relative performance evaluations may have on the average number of answers shared in Part 2 for different standardized effect sizes.[37] I do the power analysis for *standardized* effect sizes because this way, I can obtain benchmark effect sizes

majority of these missed opportunities are, indeed, missed opportunities rather than people having simply guessed the right answer to a question that they did not share because they thought they would not know the answer.

[37] The standardized effect size d is defined in the following way $d = \frac{\mu_{helpbas} - \mu_{helptreat}}{SD_{help}}$. I take this study's sample size N=282 as fixed and determine the power of a two-sided t-test to reject the null hypothesis of no effect at the 0.05 level of statistical significance or higher for different d.

from related studies, in addition to the conventional "large", "medium" and "small" effect sizes that Cohen (1977) first suggested for the social sciences.

Table 8 Power of Study for Different Standardized Effect Sizes

source	stand. effect size (d)	power
Cohen (1977)	0.8	0.998
Carpenter et. al. (2012)	0.621	0.959
Cohen (1977)	0.5	0.846
Buser and Dreber (2016)	0.287	0.401
Cohen (1977)	0.2	0.222

Notes. The reported power is the one of a two-sided t-test with a level of significance of at least α=0.05 given this study's sample size of 282 participants.

Buser and Dreber (2016) and Carpenter, Matthews and Schirm (2010) have somewhat related experimental contexts[38] and I calculated standardized effect sizes for their relevant main outcome variables and treatment comparisons (see the Online Appendix for a detailed description). Buser and Dreber (2016) find that average contributions in a Public Goods Game are lower under relative pay concerns (standardized effect size d=0.287, two-sided t-test p=0.001). Carpenter at al. (2010) find that group members provide much less favorable assessments of the quality of their peers' work output after the experience of a competition for relative pay (standardized effect size d=|0.621|[39], two-sided t-test p=0.001). My study has a power of 0.4 to detect a stand. effect size d=0.287 of relative performance concerns on average help and a power close to 1 (0.959) if the true effect size was d=0.62. For what is typically considered a medium effect size of d=0.5 my study would reject the null hypothesis of no effect 85% of the times, with the likelihood of a Type-2 error at 15%.

In other words, the results from this study tell us that it is very unlikely that relative performance concerns—caused by relative performance evaluations or relative performance pay—have a substantial negative effect on the willingness to help others perform better at work.[40]

[38] Both studies also have a between-subject design in which behavior under relative *pay* concerns (pay tournament) is compared to behavior in a baseline in which piece-rates are paid. Both studies document that competing for relative pay in a first stage has negative behavioral spillovers to subsequent prosocial or anti-social behavior in groups when the two stages are in no way strategically linked.

[39] Due to the construction of the variable, the stand. effect size is -0.621 (negative 0.621) in Carpenter et al. (2010) but with the same implication: the average assessment of other group members' production quality is much lower under the impression of relative pay concerns in reference groups.

[40] The findings of this study cannot speak as confidently about small effect sizes. Reassuringly, if we think about the policy implications of this work, it would be most important to know about sizable costs that relative performance evaluations may have on how members of a reference group work together.

The knowledge sharing task in this study models the type of on-the-job help that is important for productivity in organizations that, for example, center around knowledge work.

This experimental study can pin down the effect of relative performance evaluations on group morale because I rule out one of the major confounding factors that arise in organizational settings outside of the laboratory in which, at least in the medium run, higher relative performances may come with monetary rewards. There are several reasons why my main result is informative about how employees or students outside of the laboratory work together. Experimental reference groups in the sample of students are similar to students' actual "professional" reference groups. Moreover, knowledge-sharing is an important dimension of how co-workers in firms can help the productivity of each other.

In important ways, my experimental decision environment was conducive to uncovering a negative effect of relative performance concerns on the willingness to help others perform better. Firstly, the performance ranking was established on general knowledge and the helping behavior involved sharing knowledge in the same type of general knowledge questions. The spillover from a performance competition should be largest when relative performance rankings are established on a particular job and then co-workers can help each other to perform better on that job. Secondly, helping others in the reference group was costly and there was no monetary benefit to helping for the person that extends help. This makes helping in my experiment a generous act. Moreover, in the absolute majority of reference groups (95%), the members did not know each other prior to the experiment[41] and the experimental protocol did not reveal anything about the group members' identities other than what is revealed in a portrait. This type of intrinsically motivated help should be most sensitive to changes in the level of competition in reference groups, in particular when strangers are at the receiving end of it. Thirdly, the experimental design rules out that others can observe individual help behavior and there are no repeated interactions. Participants helped privately and made all helping decisions once and at the same time. In this environment, the decision not to help was essentially impossible to detect and so was the decision to help.

Therefore, the fact that I find no evidence that relative performance evaluations negatively affect group morale in this decision environment is particularly informative. There is no reason to believe that relative performance concerns may lower the willingness to help others perform better in many other relevant contexts, e.g. when help behavior is observable or when there are monetary rewards to it.

Previous experimental work on on-the-job help typically modeled help as the transfer of money. The task introduced in this experiment measures "real help" behavior of participants who can share knowledge for the benefit of others. The knowledge-sharing task is versatile and

[41] Out of 282 participants, 6 participants (2%) answered that they knew one person in his or her reference group and 1 participant (0.3%) answered that he or she knew both members of his or her reference group.

easy to implement in a laboratory setting. I hope that this task will prove useful for researchers who are interested in studying the determinants of on-the-job help.

There are now several studies, including this one, that come to the conclusion that relative performance ranks make a qualitative difference compared to an environment in which this information is not given but not whether it is provided in public or private (see also Tran and Zeckhauser 2012; Ashraf et al. 2014). A general lesson for feedback design seems to emerge here, namely, that relative performance information, provided in private or in public, will put members of the reference group in a comparative and also competitive mind-set.

My results provide first evidence on how the intensity of competition in reference groups compares under relative performance information and relative pay. There is no evidence that the introduction of relative pay further increases the level of competition in reference groups. This finding suggests that, at least in some contexts, relative performance evaluations may substitute relative pay in tournaments. An interesting avenue for future research is to systematically compare individual effort choices under non-monetary and monetary tournaments and when they interact.

My main result on the willingness to help across experimental conditions advances our understanding of when relative performance evaluations do, and when they do not backfire in reference groups. Charness, Masclet and Villeval (2014) find that relative performance rankings lead to costly unethical behavior in reference groups because group members want to change an initial ranking to a final ranking that is more favorable for them. The authors interpret this main finding as an expression of group members' competitive preferences and their desire for dominance. The decision environment of this study completely removed the strategic link between knowledge-sharing and the ranking outcome itself and competitive preferences under relative performance evaluations did not have a negative consequence on help behavior. Taken together, these results suggest that the frequency of relative performance evaluations may determine whether ore not they have a negative impact on the way members of reference group work together.

The result that there is no evidence of a negative spillover of the competition for relative ranks or for relative pay on group morale is surprising. It may be due to the fact that the ranking was an objective assessment of relative performances. Group members knew on what grounds the ranking was established and that the computer impartially implemented the ranking based on performance at the first general knowledge test. The findings by Breza et al. (2018) lend support to this speculation. They find that the negative effects of relative performance pay in workgroups depend on how transparent it was to co-workers that others in the reference group were more productive. Their results indicate that co-workers who knew that pay differences arose from observable performance differences did not react negatively, either by exerting less effort or cooperating less well. In organizations, relative performance evaluations can also mirror

subjective perceptions of managers. At least employees can perceive this to be the case. It may, therefore, be advisable to transparently communicate criteria and performance metrics on which grounds relative performance evaluations are established. It would be interesting to investigate in future work to what extent subjectivity and transparency of evaluation criteria mediate the effect that relative performance evaluations have on group morale or other workgroup behavior.

My results thus draw attention to the challenge of understanding better when relative concerns among employees do and when they do not backfire in reference groups. I find no evidence that relative performance evaluations, whether or not they entail monetary consequences, may have negative consequences for group productivity. The unambiguous positive lesson from this study is that there is no evidence that relative performance concerns in reference groups lower the intrinsic motivation to help others perform better.

References

Ashraf, Nava, Oriana Bandiera, and Scott S. Lee. 2014. "Awards Unbundled: Evidence from a Natural Field Experiment." *Journal of Economic Behavior & Organization* 100 (April): 44–63. https://doi.org/10.1016/j.jebo.2014.01.001.

Austen-Smith, David, and Roland G. Fryer Jr. 2005. "An Economic Analysis of 'Acting White.'" *The Quarterly Journal of Economics* 120 (2): 551–583.

Bandiera, Oriana, Iwan Barankay, and Imran Rasul. 2013. "Team Incentives: Evidence from a Firm Level Experiment." *Journal of the European Economic Association* 11 (5): 1079–1114. https://doi.org/10.1111/jeea.12028.

Bénabou, Roland, and Jean Tirole. 2006. "Incentives and Prosocial Behavior." *American Economic Review* 96 (5): 1652–78. https://doi.org/10.1257/aer.96.5.1652.

Besley, Timothy, and Maitreesh Ghatak. 2008. "Status Incentives." *The American Economic Review* 98 (2): 206–211.

Black, Paul W, Andrew H. Newman, Bryan Stikeleather, and Nathan J. Waddoups. 2018. "Performance Feedback Type and Employees Subsequent Willingness to Help Other Employees." *Journal of Management Accounting Research*, October. https://doi.org/10.2308/jmar-52298.

Blanes i Vidal, Jordi, and Mareike Nossol. 2011. "Tournaments Without Prizes: Evidence from Personnel Records." *Management Science* 57 (10): 1721–36. https://doi.org/10.1287/mnsc.1110.1383.

Breza, Emily, Supreet Kaur, and Yogita Shamdasani. 2018. "The Morale Effects of Pay Inequality." *The Quarterly Journal of Economics* 133 (2): 611–63. https://doi.org/10.1093/qje/qjx041.

Bursztyn, Leonardo, Thomas Fujiwara, and Amanda Pallais. 2017. "'Acting Wife': Marriage Market Incentives and Labor Market Investments." *American Economic Review* 107 (11): 3288–3319. https://doi.org/10.1257/aer.20170029.

Buser, Thomas, and Anna Dreber. 2016. "The Flipside of Comparative Payment Schemes." *Management Science* 62 (9): 2626–38. https://doi.org/10.1287/mnsc.2015.2257.

Carpenter, Jeffrey, Peter Hans Matthews, and John Schirm. 2010. "Tournaments and Office Politics: Evidence from a Real Effort Experiment." *American Economic Review* 100 (1): 504–17. https://doi.org/10.1257/aer.100.1.504.

Charness, Gary, Uri Gneezy, and Austin Henderson. 2018. "Experimental Methods: Measuring Effort in Economics Experiments." *Journal of Economic Behavior & Organization* 149 (May): 74–87. https://doi.org/10.1016/j.jebo.2018.02.024.

Charness, Gary, David Masclet, and Marie Claire Villeval. 2014. "The Dark Side of Competition for Status." *Management Science* 60 (1): 38–55. https://doi.org/10.1287/mnsc.2013.1747.

Coffman, Katherine Baldiga. 2014. "Evidence on Self-Stereotyping and the Contribution of Ideas *." *The Quarterly Journal of Economics* 129 (4): 1625–60. https://doi.org/10.1093/qje/qju023.

Cohen, Jacob. 1977. "CHAPTER 2 - The t Test for Means." In *Statistical Power Analysis for the Behavioral Sciences (Revised Edition)*, 19–74. Academic Press. https://doi.org/10.1016/B978-0-12-179060-8.50007-4.

Danilov, Anasatsia, Christine Harbring, and Bernd Irlenbusch. 2014. "Helping in Teams." https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2543901.

Deming, David J. 2017. "The Growing Importance of Social Skills in the Labor Market." *The Quarterly Journal of Economics* 132 (4): 1593–1640. https://doi.org/10.1093/qje/qjx022.

Drago, Robert, and Gerald T. Garvey. 1998. "Incentives for Helping on the Job: Theory and Evidence." *Journal of Labor Economics* 16 (1): 1–25. https://doi.org/10.1086/209880.

Duffy, John, and Tatiana Kornienko. 2010. "Does Competition Affect Giving?" *Journal of Economic Behavior & Organization* 74 (1–2): 82–103. https://doi.org/10.1016/j.jebo.2010.02.001.

Ellingsen, Tore, and Magnus Johannesson. 2008. "Pride and Prejudice: The Human Side of Incentive Theory." *American Economic Review* 98 (3): 990–1008. https://doi.org/10.1257/aer.98.3.990.

Erkal, Nisvan, Lata Gangadharan, and Nikos Nikiforakis. 2011. "Relative Earnings and Giving in a Real-Effort Experiment." *The American Economic Review* 101 (7): 3330–48.

Ewers, Mara, and Florian Zimmermann. 2015. “IMAGE AND MISREPORTING: Image and Misreporting.” *Journal of the European Economic Association* 13 (2): 363–80. https://doi.org/10.1111/jeea.12128.

Falk, Armin, Anke Becker, Thomas J. Dohmen, David Huffman, and Uwe Sunde. 2016. “The Preference Survey Module: A Validated Instrument for Measuring Risk, Time, and Social Preferences.” https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2725035.

Fischbacher, Urs. 2007. “Z-Tree: Zurich Toolbox for Ready-Made Economic Experiments.” *Experimental Economics* 10 (2): 171–78. https://doi.org/10.1007/s10683-006-9159-4.

Gächter, Simon, Chris Starmer, and Fabio Tufano. 2015a. “Measuring the Closeness of Relationships: A Comprehensive Evaluation of the ‘Inclusion of the Other in the Self’ Scale.” *PLOS ONE* 10 (6): e0129478. https://doi.org/10.1371/journal.pone.0129478.

Gächter, Simon, Christian Starmer, and Fabio Tufano. 2015b. “Measuring the Impact of Social Relationships: The Value of ‘Oneness.’” Mimeo.

Gill, David, Zdenka Kissová, Jaesun Lee, and Victoria Prowse. 2018. “First-Place Loving and Last-Place Loathing: How Rank in the Distribution of Performance Affects Effort Provision.” *Management Science*, February. https://doi.org/10.1287/mnsc.2017.2907.

Karni, Edi. 2009. “A Mechanism for Eliciting Probabilities.” *Econometrica* 77 (2): 603–6. https://doi.org/10.3982/ECTA7833.

Köszegi, Botond. 2006. “Ego Utility, Overconfidence, and Task Choice.” *Journal of the European Economic Association* 4 (4): 673–707. https://doi.org/10.1162/JEEA.2006.4.4.673.

Kuhnen, Camelia M., and Agnieszka Tymula. 2012. “Feedback, Self-Esteem, and Performance in Organizations.” *Management Science* 58 (1): 94–113. https://doi.org/10.1287/mnsc.1110.1379.

Lazear, Edward P. 1989. “Pay Equality and Industrial Politics.” *Journal of Political Economy* 97 (3): 561–80. https://doi.org/10.1086/261616.

Lazear, Edward P., and Kathryn L. Shaw. 2007. “Personnel Economics: The Economist’s View of Human Resources.” *Journal of Economic Perspectives* 21 (4): 91–114. https://doi.org/10.1257/jep.21.4.91.

Levine, David K. 1998. “Modeling Altruism and Spitefulness in Experiments.” *Review of Economic Dynamics* 1 (3): 593–622. https://doi.org/10.1006/redy.1998.0023.

McManus, T. Clay, and Justin M. Rao. 2015. “Signaling Smarts? Revealed Preferences for Self and Social Perceptions of Intelligence.” *Journal of Economic Behavior & Organization* 110 (February): 106–18. https://doi.org/10.1016/j.jebo.2014.12.009.

Rabin, Matthew. 1993. “Incorporating Fairness into Game Theory and Economics.” *The American Economic Review* 83 (5): 1281–1302.

Tran, Anh, and Richard Zeckhauser. 2012. “Rank as an Inherent Incentive: Evidence from a Field Experiment.” *Journal of Public Economics* 96 (9–10): 645–50. https://doi.org/10.1016/j.jpubeco.2012.05.004.

Appendix

The Appendix presents the following additional figures:

1. figures of the feedback screens in the baseline, the private rank feedback condition and the public rank feedback condition
2. pictogram of the Inclusion of the Self in Other (IOS) scale
3. empirical distribution of performance on the timed general knowledge test by experimental condition
4. figures of participants' evaluations of the decision environment in Part 1

The Appendix contains the following additional tables:

1. Results of OLS regressions that show how a participant's confidence in his group members ability to provide correct answers to Part 2 questions changes when relative performance feedback is given. This analysis conditions on the rank of the participant who makes the judgements.
2. List and summary statistics of all questionnaire items from which the indices on attitudes toward cooperation, competition, group work and autonomy are constructed
3. Rank Feedback, Relative Pay and Expected Help, General Cooperativeness and Pro-social Inclinations
4. Results of OLS regressions and Probit models predicting the willingness to share answers with others comparing the baseline to the relative pay condition, controlling for beliefs about correct answers
5. A table that shows the average efficiency loss in groups due to missed opportunities to help and the unrealized gains in group productivity due to missed opportunities to help by experimental conditions

Figure A1. Information Provided at the End of Part 1 (Baseline Condition)

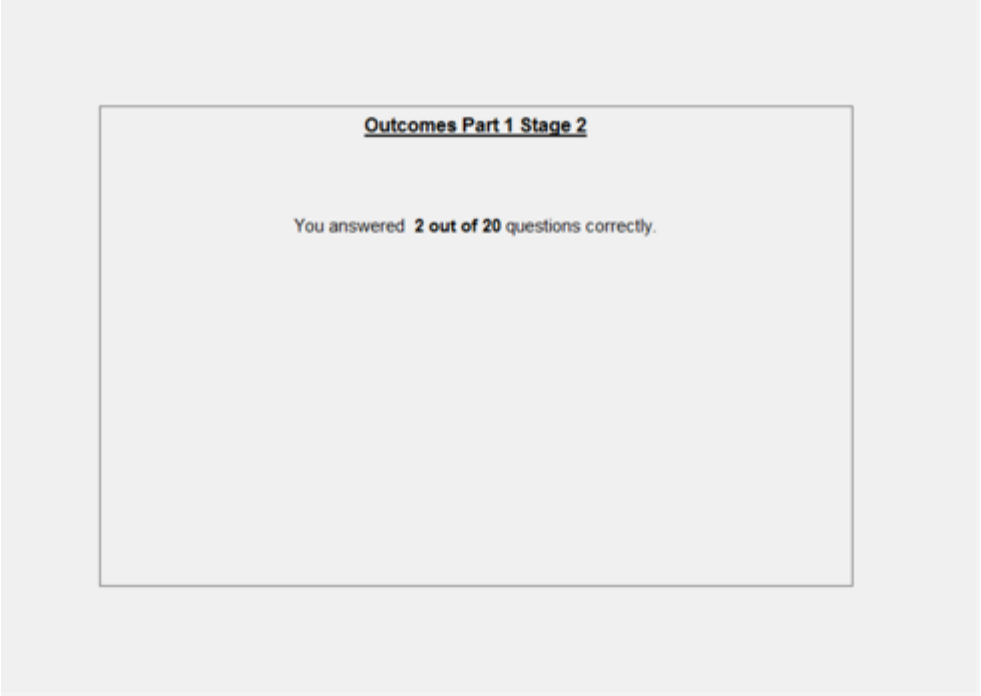


Figure A2. Information Provided at the End of Part 1 (Private Rank Feedback Condition)

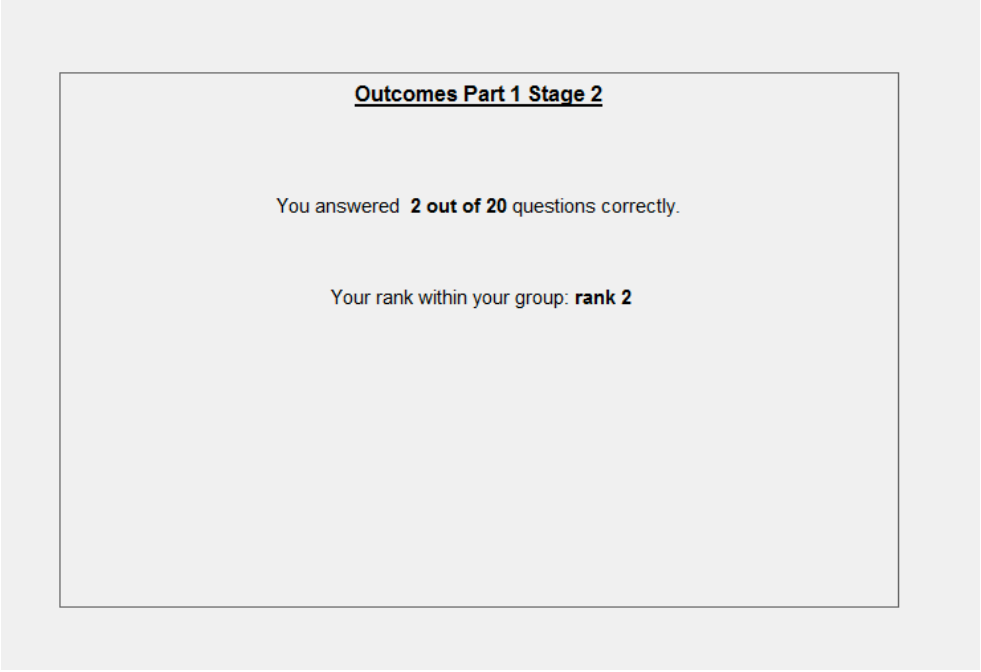
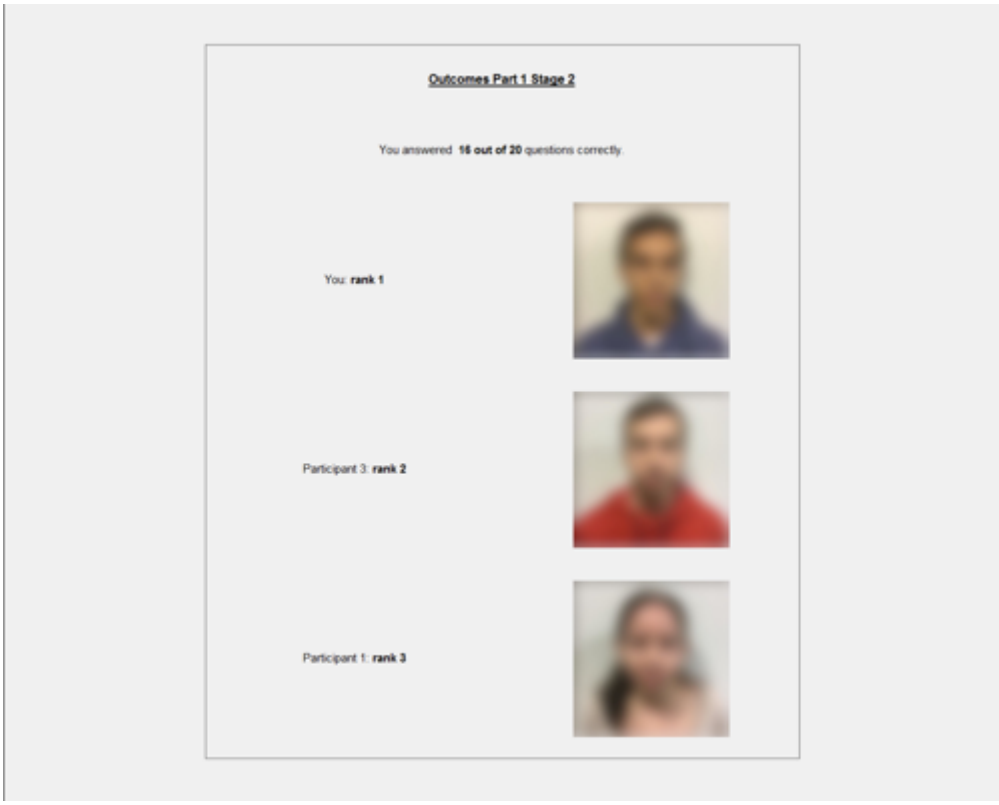


Figure A3. Information Provided at the End of Part 1 (Public Rank Feedback Condition)



Notes. The faces of participants are blurred here to preserve their anonymity.

Figure A4. Information Provided at the End of Part 1 (Relative Pay Condition)

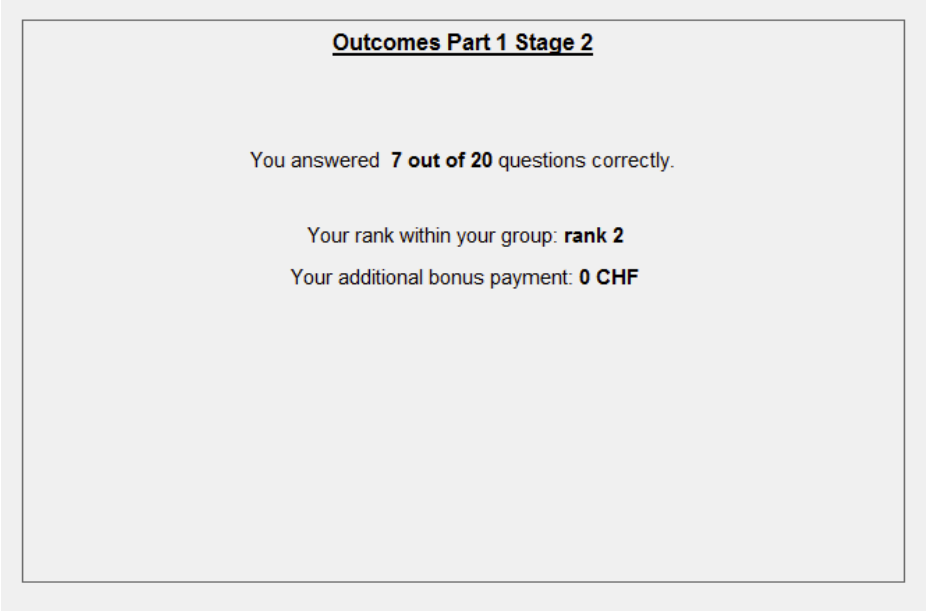


Figure A5. Pictogram of the Inclusion of the Self in Other (IOS) scale

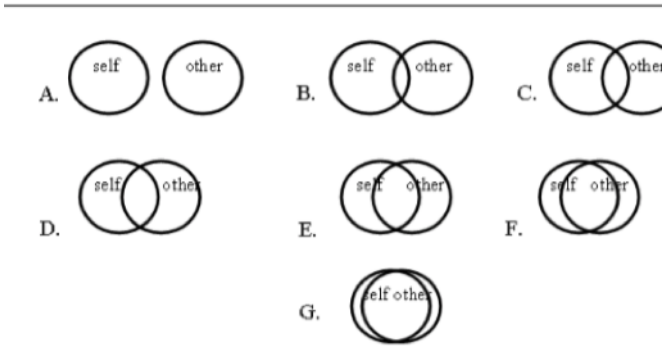


Figure A6. Empirical CDF of Performance on Timed General Knowledge Test by Condition

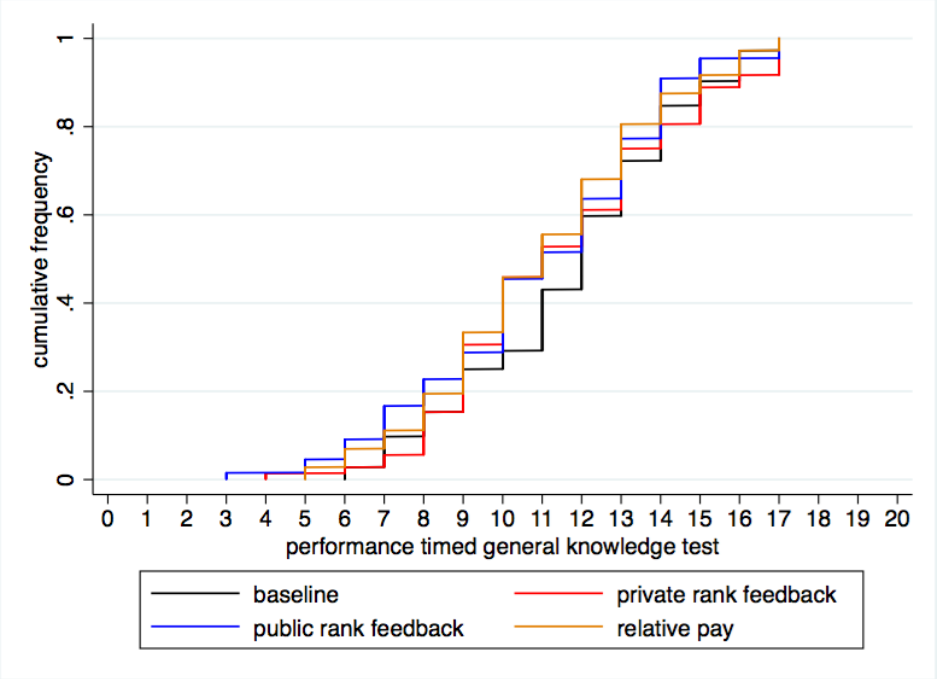


Figure A7. Participants' Evaluations of Part 1 Decision Environment

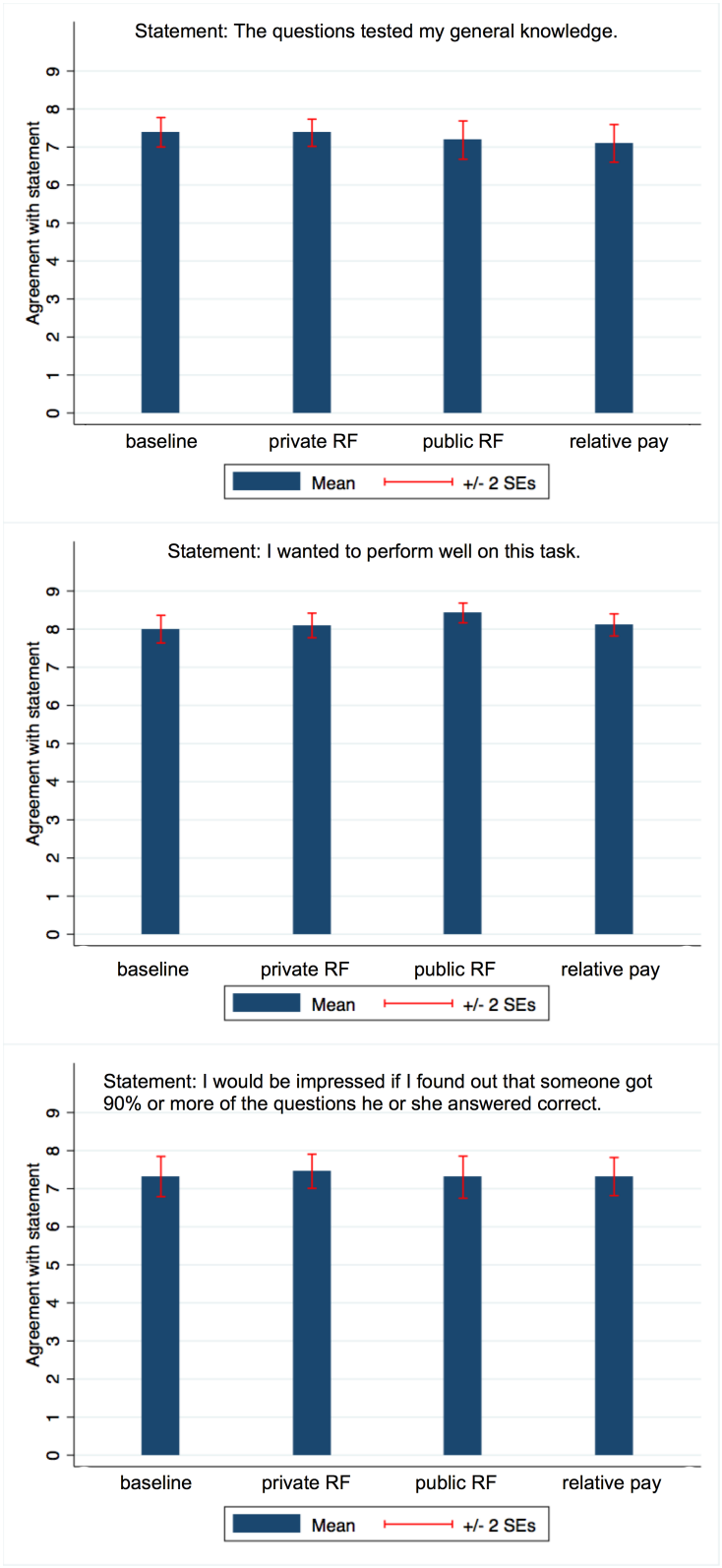


Table A1. Rank Feedback and Confidence in Others' Performance

OLS predicting belief of participant i of per cent chance that group member (GM) j answered question k correctly							
	Rank 1 participant *i*			Rank 2 participant *i*		Rank 3 participant *i*	
Private RF	-5.546**	-5.262*	Private RF	-0.606	-1.201	10.09****	10.07****
	(2.755)	(2.698)		(3.193)	(3.227)	(2.948)	(2.893)
Public RF	-7.171***	-7.358***	Public RF	2.348	-0.943	11.49****	9.153**
	(2.702)	(2.551)		(3.425)	(3.665)	(3.421)	(3.481)
GM *j* rank 3		-0.830	GM *j* rank 1		0.0935		-0.153
		(1.227)			(1.512)		(0.835)
Priv. RF X		-0.571	Priv. RF X		1.187		-0.0897
GM *j* rank 3		(1.441)	GM *j* rank 1		(1.865)		(1.158)
Pub. RF X		0.362	Pub. RF X		6.518****		4.558**
GM *j* rank 3		(2.072)	GM *j* rank 1		(1.864)		(1.759)
Controls	*Yes*	*Yes*		*Yes*	*Yes*	*Yes*	*Yes*
Constant	61.44****	61.83****	Constant	25.66***	25.50***	41.81****	41.76****
	(12.04)	(12.04)		(8.827)	(8.788)	(8.193)	(8.261)
Obs.	1400	1400	Obs.	1400	1400	1400	1400
R^2	0.1391	0.1396		0.1585	0.1636	0.1327	0.1354

Notes. Private RF and *Public RF* are indicators for participant i privately observing his performance rank on the timed task or publicly observing the performance rank of everyone in his group, respectively. *GM j rank 1* and *GM j rank 3* are indicators for whether the group member that participant i judges has performance rank 1 or 3, respectively, on the timed task in Part 1. These indicators are also interacted with the treatment indicators. The following variables are controlled for. *Actual correct* variable indicates whether *participant i* or the *group member j t*hat he judges provided a correct answer to a Part 2 question k. *Performance Part 1* records the number of questions that participant i answered correctly during the timed task in Part 1. *Risk attitude* is where participant i positioned himself on a scale from 0=very risk-averse to 10=very risk seeking. *Female* indicates whether participant i is a woman. *Female group member* indicates whether group member j is a woman. *Significant at the 10% level,** at the 5% level, *** at the 1% level, **** at the 0.1% level.

Table A2. Questionnaire Items on Attitudes Toward Cooperation, Competition, Working in Groups and Working Alone

Item	Category	Mean	SD
I am drawn to compete with others.	Competitiveness	4.98	2.07
It annoys me when others perform better than I do.	Competitiveness	5.16	2.15
I feel that winning or losing doesn't matter to me.	Competitiveness (-)	3.86	2.09
I avoid competitive situations.	Competitiveness (-)	4.57	2.24
I love to help others.	Cooperativeness	7.33	1.34
I like to share my ideas and material with others.	Cooperativeness	6.54	1.63
I avoid doing favors to others.	Cooperativeness (-)	2.70	1.73
I expect everyone to look out for themselves.	Cooperativeness (-)	5.57	1.94
I like to work things out on my own.	Autonomy	6.92	1.63
Given the choice, I prefer to work on an assignment alone rather than getting an assignment in which I have to work together with others.	Autonomy	5.37	2.14
I find it hard to work by myself.	Autonomy (-)	2.73	1.55
I find I am less productive when I work by myself.	Autonomy(-)	3.05	1.83
I can learn important things from other colleagues or fellow students.	Groupwork	7.94	1.29
I like working in groups.	Groupwork	5.79	1.97
In workgroups, one person does typically most of the work.	Groupwork (-)	5.88	1.94
I find that working in groups is often inefficient.	Groupwork (-)	5.65	1.97

Notes. This table summarizes all the fours items from which the index for that category is constructed. The answers to each question ranges from 1- does not apply at all to 9- definitely applies. The index is the average score across the four items of a category and negatively keyed items are reverse scored.

Table A3. Rank Feedback, Relative Pay and Expected Help, General Cooperativeness and Prosocial Inclinations

	Expected Help	Cooperativeness	Positive Reciprocity	Negative Reciprocity
Private RF	-0.201	-0.101	-1.008*	-0.771
	(0.389)	(0.175)	(0.522)	(0.744)
Public RF	-0.0789	-0.0571	-0.457	0.00184
	(0.388)	(0.176)	(0.556)	(0.777)
Relative Pay	-0.403	0.0590	-0.485	-0.522
	(0.417)	(0.211)	(0.577)	(0.874)
Constant	4.632****	5.955****	13.63****	10.83****
	(0.215)	(0.124)	(0.391)	(0.513)
Obs.	282	282	282	282
R^2	0.0039	0.0030	0.0093	0.0048

Notes. Private RF and *Public RF* are indicators for participant i privately observing his performance rank on the timed task or publicly observing the performance rank of everyone in his group, respectively. *Expected help* is the number of answers to Part 2 questions that participant i thinks his two group members shared with him on average. Cooperativeness is an index that measures the desire to help others in general and runs from 1 (very low) to 9 (very high). Robust standard errors are in parentheses, 70 group clusters allow for correlated observations at the group and at the subject level. *Significant at the 10% level,** at the 5% level, *** at the 1% level.

Table A4. Predicting the Willingness to Help Under Relative Pay Compared to Baseline

	Predicting Prob(share answer to question k)			
	Probit		OLS	
Relative Pay	-0.0641	-0.0559	-0.06388	-0.05258
	(0.0478)	(0.0527)	(0.04784)	(0.04883)
Belief correct (self)		0.6442****		0.60298****
		(0.0950)		(0.08113)
Belief correct (others)		-0.2064*		-0.19484*
		(0.1076)		(0.09987)
Constant	Yes	Yes	0.40277****	0.1193**
			(0.0261)	(0.0522)
Obs.	1,440	1,440	1,440	1,440
(pseudo) R^2	0.003	0.088	0.004	0.1109

Notes. Predicted marginal effects at mean level of covariates are reported (Probit). *Share answer to question k* is an indicator for whether participant shared the answer to a Part 2 question k with others. *Relative Pay* is an indicator for participant i privately observing his performance rank on the timed task and the best performer receiving a substantial monetary bonus at the end of the timed task. *Belief correct self* ranges from [0,1] and is the subjective probability that participant i gives to the event that his answer to question k is correct. *Belief correct others* ranges from [0,1] and is the subjective probability that participant i gives to the event that his average group member provided the correct answer to question k. *Significant at the 10% level,** at the 5% level, *** at the 1% level, **** at the 0.1% level.

Table A5. Missed Opportunities to Help, Efficiency and Group Productivity by Experimental Condition

	Average efficiency loss	Average gains in group productivity
Baseline	5.12 CHF	15.2%
Private RF	5.29 CHF	16.3%
Public RF	6.57 CHF	21.7%
Relative Pay	6.27 CHF	20.0%
Total	5.80 CHF	18.2%

Notes. Average efficiency loss due to missed opportunities to help is the amount of unrealized group earnings net the cost of help in a group, averaged over all groups. Average gains in group productivity is the number of missed opportunities in a group over a group's total productivity, averaged over all groups.