European Economic Review 156 (2023) 104458

Contents lists available at ScienceDirect

European Economic Review

journal homepage: www.elsevier.com/locate/eer

Monetary and social incentives in multi-tasking: The ranking substitution effect☆

Matthias Stefan [a], Jürgen Huber [a], Michael Kirchler [a,*], Matthias Sutter [b,c,d], Markus Walzl [e]

[a] *University of Innsbruck, Department of Banking and Finance, Universitätsstrasse 15, Innsbruck, Austria*
[b] *Max Planck Institute for Research on Collective Goods, Kurt-Schumacher-Strasse 10, Bonn, Germany*
[c] *University of Cologne, Department of Economics, Albertus-Magnus-Platz, Cologne, Germany*
[d] *University of Innsbruck, Department of Public Economics, Innsbruck, Austria*
[e] *University of Innsbruck, Department of Economics, Universitätsstrasse 15, Innsbruck, Austria*

A R T I C L E I N F O

Dataset link: https://osf.io/ujncd/

JEL classification:
C93
D02
D91

Keywords:
Multi-tasking decision
Rank incentives
Artefactual field experiment
Finance professionals

A B S T R A C T

Rankings are intended as incentive tools on labor markets. Yet, when agents perform multiple tasks, rankings might have unintended side-effects, especially if not all tasks can be ranked with respect to performance. We analyze the dynamics of multi-tasking and present an experiment with 286 finance professionals in which we identify hidden ranking costs when performance in one task is ranked while in another prosocial task it is not. We find that subjects lagging behind (leading) in the ranked task devote less (more) effort to the prosocial task. We discuss implications for optimal incentive schemes in organizations with multi-tasking.

1. Introduction

Monetary incentives and (non-pecuniary) social comparisons like rank incentives are among the most prevalent incentive structures in companies and public institutions (Coles et al., 2018). Examples include combinations of fixed salaries and bonus payments, relative-performance evaluations (tournaments), but also institutional designs that make use of social (peer) comparison to promote employee performance. However, numerous incidences of moral hazard, outright fraud, and bad performance indicate that ill-designed incentive structures can lead to antisocial behavior and negative externalities for the company and society as a whole. This is particularly important in cases of multi-tasking when a prosocial component is involved, as relative-performance evaluations of the non-prosocial task might potentially crowd out prosocial behavior and thus lead to efficiency losses for the

☆ Our experimental software, anonymized data files, and the analysis script are stored on the OSF repository https://osf.io/ujncd/.

We thank Felix Holzmeister, Christian König, Nora Szech and seminar and conference participants at the SFB "Credence Goods, Incentives, and Behavior"-Workshop 2019 Innsbruck and the ZEW Workshop on Market Design 2019 for very valuable comments on earlier versions of this paper. Financial support from the Austrian Science Fund (FWF START-grant Y617-G11, SFB F63) and the Swedish Research Council (grant 2015-01713) is gratefully acknowledged. This study was ethically approved by the IRB of the University of Innsbruck.

* Corresponding author.

E-mail addresses: matthias.stefan@uibk.ac.at (M. Stefan), juergen.huber@uibk.ac.at (J. Huber), michael.kirchler@uibk.ac.at (M. Kirchler), matthias.sutter@coll.mpg.de (M. Sutter), markus.walzl@uibk.ac.at (M. Walzl).

https://doi.org/10.1016/j.euroecorev.2023.104458
Received 5 October 2021; Received in revised form 12 March 2023; Accepted 8 April 2023
Available online 11 May 2023
0014-2921/

company and society at large. Surprisingly, little scientific evidence exists on trade-off decisions between individual monetary and rank incentives on the one hand and prosocial behavior on the other hand.

In this paper, we narrow this research gap and investigate how monetary incentives and social comparison influence effort in a multi-tasking environment with a trade-off between one's own payment/rank and a prosocial activity. First, we analyze decision making of an agent who receives monetary payments for one task, but may also be intrinsically motivated to spend effort on a prosocial task, thus facing a trade-off decision.[1] As a variation we introduce a ranking on the monetarily incentivized task, which informs the agent about her performance in this task relative to other agents. Depending on how the agent interprets the ranking, it can work as an additional incentive or disincentive for the monetarily incentivized task: If a good rank is desirable, it can increase total effort and the fraction of effort spent for the ranked task. If a good rank is costly, because it is regarded as a signal (to the self or others) of low prosocial activity, it might lower total effort and the fraction of effort spent on the ranked task. Following the literature in social psychology (see e.g., Fishbach et al., 2009), one could expect the agent to balance (or highlight, i.e., focus on one of) the conflicting motives over time. For balancing, we expect a ranking to be a disincentive for the ranked task if the agent's performance was ranked highly in previous periods and an incentive for the ranked task in case of a poor ranking in previous periods.

Second, we test the hypotheses derived in this framework in a controlled online experiment with 286 internationally operating finance professionals (e.g., fund managers, private bankers, traders). They represent a suitable subject pool as they are exposed to various forms of social comparison in their profession and regularly face trade-offs between monetary returns and prosocial decision making, e.g., in the context of sustainable investments. The professionals had to solve items of an IQ-related test – i.e., the Raven's Advanced Progressive Matrices test (Raven, 2000) – and had to choose whether to solve these items for their own payment (the selfish activity) or for donations for measles vaccine to UNICEF (the prosocial activity). In a 2 × 3 factorial experimental design, we varied the private display of one's individual rank among peers (displayed or not) and the level of monetary incentives (high, medium and low piece-rates) where both incentives are only relevant for the items solved for oneself and not for the items solved for donation.

As our main contribution, we show that the introduction of a ranking on the monetarily incentivized activity leads to a *ranking substitution effect*: Those professionals that are ahead in the ranking substitute relative effort spent for their own payment by putting more effort into the prosocial activity — i.e., the impact of a ranking is similar to a reduced piece-rate. In contrast, those lagging behind substitute by spending more effort for their own payment and less for the prosocial activity — i.e., the ranking works like an additional piece-rate. Hence, the benefits of a ranking come at a cost: While some individuals act more prosocially, others focus more on the payoff-relevant task. We discuss our results in light of the optimal design of incentive schemes and labor market contract regulations to overcome moral hazard and adverse selection in (imperfectly competitive) labor markets for managerial talent (Bénabou and Tirole, 2016).

Our study is motivated by real-world anecdotal evidence on the role of monetary incentives and social comparison. To name just two examples: First, there has been a widespread debate on the "Wall Street culture" of incentive schemes in the finance industry. While these incentive schemes are pecuniary by nature, they also include a strong element of social comparison of salary among peers. In some sub-sectors like the (hedge) fund industry, professionals' salaries are a convex function of past performance relative to other fund managers, resulting in a tournament incentive structure (Brown et al., 1996; Sirri and Tufano, 1998; Kaniel and Parham, 2017). These incentive schemes, however, have been criticized as one of the potential drivers of excessive risk taking in the finance industry (Rajan, 2006; Diamond and Rajan, 2009; Kirchler et al., 2018). This debate reflects public worries about the potential detrimental effects of monetary and rank incentives "gone wild" by generating negative externalities. It also shows that prosocial behavior, such as contributing to a public good like financial stability, can potentially be overruled by the individual aspiration for high social status and monetary payments.

Second, publication merits in academia as a form of social comparison are not directly related with researchers' monetary incentives. While it is true that the scientific reputation of some positions and their payment are positively correlated, the hunt for top-ranked publications can hardly be explained solely by monetary incentives, in particular not for tenured senior researchers. As in the finance industry, it has been argued that status seeking may crowd-out intrinsic research motivation (Osterloh and Frey, 2015) and may be one of the reasons for misconduct and sabotage among researchers (Anderson et al., 2007; Fanelli, 2010). Similar to the set-up we examine, researchers may have to trade off between high publication reputation and contributing to the public good of service to the community and students.

Our paper particularly contributes to literature on social comparison (Festinger, 1954; Bandiera et al., 2010; Cohn et al., 2015) and status (Moldovanu et al., 2007). Various studies disentangling rank incentives (social comparison) from monetary incentives show the effect of non-incentivized rankings on performance (Tran and Zeckhauser, 2012; Barankay, 2015), portfolio choice (Dijk et al., 2014), risk taking (Kuziemko et al., 2014; Kirchler et al., 2018), and market prices (Ball et al., 2001). With regards to effort provision, the literature reports varying effects of rankings, ranging from an overall increase in effort (Azmat and Iriberri, 2010; Blanes-i-Vidal and Nossol, 2011; Tran and Zeckhauser, 2012) to effects depending on, for instance, expectations, current rank, details of the principal agent relationship and gender (Al-Ubaydli and List, 2015; Kuhnen and Tymula, 2012; Gill et al., 2019; Murad et al., 2019).[2] On a more general level, our paper also contributes to the literature on the relation between incentive

[1] Similar to Bénabou and Tirole (2016), one could interpret the task as observable but not verifiable to the employer or the company.

[2] Moreover, peer effects need not necessarily be part of an explicit incentive scheme or an explicitly designed ranking, but can also emerge rather naturally (Mas and Moretti, 2009).

schemes and performance measures (e.g., Baker, 1992; Bénabou and Tirole, 2016). Our inquiry is also related to studies on self-image concerns (Bénabou and Tirole, 2006), which, by their very nature, crucially depend on the specific set-up (Ariely et al., 2009; Falk and Szech, 2019). From this perspective, the trade-off choice in our model can also be interpreted as a balancing of the desire for a positive self-image due to observed rank and the desire for a positive self-image that stems from contributing to a prosocial activity. This also relates to the literature discussing how social comparison and monetary incentives can lower prosocial behavior or even promote misconduct (Shleifer, 2004; Charness et al., 2014).

While existing studies mainly focus on effort in one domain, we extend the literature by studying a dynamic multi-tasking problem where subjects can distribute their effort between a selfish and a prosocial activity over time.[3] The novel aspect of our design is that subjects can distribute effort between two activities (i.e., a feature of many real-world decisions) allowing us to separately analyze implications for total effort and for substitution effects between both activities.

2. The experiment

2.1. Experimental design and treatments

In our online experiment, subjects had to solve items of the Raven's Advanced Progressive Matrices test (APM; Raven, 2000). In each of the APM items, subjects had to recognize the geometric pattern in an unfinished diagrammatic puzzle and identify the missing element. The main objective of the APM is to measure subjects' ability to solve novel problems, which is why it is also used as a measure of IQ (see Fig. 1 for an example of an APM item and the full experimental instructions in Section C in the Appendix). One advantage of APM is that subjects are expected to be intrinsically motivated to obtain a higher rank in an intelligence test (Falk and Szech, 2019). Moreover, evidence shows that performance in IQ tests and related tasks are not only a measure of ability but also effort and do indeed respond to incentives (Borghans et al., 2013; Gneezy et al., 2019). Finally, there are no expected learning effects in APM (Lozano and Revuelta, 2020).

Subjects participated in four periods of two minutes each, consisting of two treatments (see details below). The order of items was randomized as follows: First, we randomized the APM items into two sequences which were then used for all participants. Then, for each subject a random draw decided which sequence occurred in which block.

In order to achieve a fair comparison, we applied the same two random sequences to all subjects and only randomly varied their ordering between blocks with and without ranking. In our experiment, participants could provide effort in two domains. Domain A, i.e., the prosocial activity, was not incentivized for the subject, but associated with a positive externality that is expected to generate an intrinsic prosocial motivation. In particular, we donated € 10 to UNICEF for measles vaccination for each solved item for *Domain A*, which was described in detail in the instructions and was therefore public knowledge (see Fig. 1 for a screenshot of the instructions outlining the prosocial activity). The choice of externality is based on Kirchler et al. (2016). In contrast to our experiment, in real-life examples "prosocial" behavior often affects the firm or co-workers and, thus, indirectly also the agent. However, in many of the prominent examples of public debates of the finance industry, the repercussion of an agent's action on herself is negligible. We would like to mention two examples: First, it is assumed that excessive risk taking during the financial crisis of 2008/2009 had harmful effects on the public, but due to option-like incentives little negative effect for many of the decision makers (Rajan, 2006; Diamond and Rajan, 2009; Bebchuk and Spamann, 2010). Second, sustainable banking can have lasting positive effects on the public, while the individual decision maker in a bank typically is only partly and indirectly affected. While there are a number of examples from real life that seem to be aligned with the externality in our experiment, the main rationale for our design choice is that we are able to cleanly separate egoistic from prosocial motives in our controlled experimental setting. For *Domain B*, in contrast, each subject received a certain piece-rate for each solved item for individual payment, which varied across treatments (€ 5, € 10, or € 15). Subjects had to decide for each APM item whether to solve it for Domain A or B *before* it was shown. Due to the random sequence, subjects could not infer the difficulty of the next item to be solved.

We set up a 2 × 3 factorial treatment design. Because of the large number of treatments for an artefactual field experiment (Harrison and List, 2004), we implemented a within-between-subjects design (see Table 1 for details).

As first treatment variation, we modified whether a ranking was shown to the participants or not. We implemented this treatment variation as a within-subject design, i.e., each subject participated in a block of two periods with and in a block of two periods without a ranking in randomized order. The individual rank was displayed at the end of each of the two periods within the block of treatments with ranking, based on a subject's total number of correctly solved APM items for Domain B within two minutes. In particular, we showed the individual rank among 13 peers (12 plus the respective subject), but no further details about their performance or the rank of other participants. Given the possibilities of ties, participants could achieve ranks $\text{RANK} \in \{1, 2, 7, 9, 12\}$. We pre-sampled the 12 professionals from the same subject pool in a first wave to constitute the ranking. They were invited to participate in an experiment that was identical to our main experiment (i.e., including monetary incentives) except for two variations: First, there were only two periods, both without a ranking; second, subjects solved the items only for Domain B. This design choice avoids that participants might learn about social norms or the "relative return to effort" by observing the ranking. Basing the ranking on pre-sampled performance allows to run an online experiment as not all subjects have to participate in the experiment at the same time (for a similar design see Kirchler et al., 2018). We deliberately chose this design to implement the mildest form of social

[3] For instance, literature on crowding-out of motivation by extrinsic incentives, too, mainly focuses on effort in one domain where subjects are (initially) intrinsically motivated (see e.g., Gneezy and Rustichini, 2000a,b; Mellström and Johannesson, 2008; Bowles and Polania-Reyes, 2012).

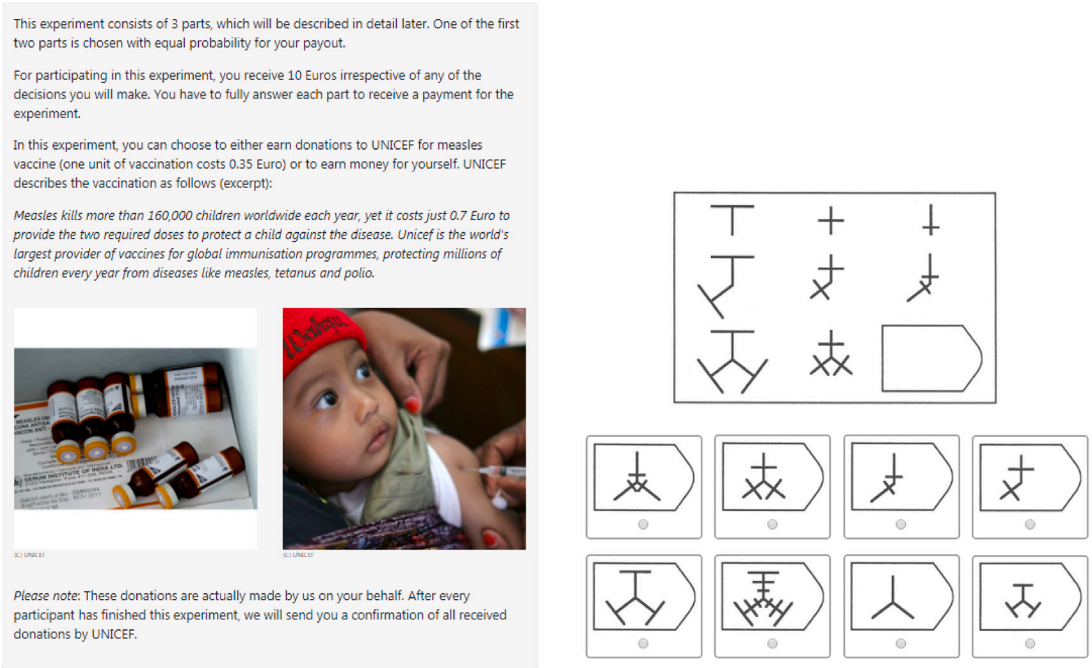

Fig. 1. Left panel: Screenshot of the instruction screen outlining the explanation of the donation to measles vaccination to UNICEF (i.e., the prosocial activity). Right panel: Screenshot example of one APM item.

Table 1
Treatment overview: "PR" (piece-rate) denotes the individual payment for each solved item in Domain B. The numbers after "PR" indicate the level of the piece-rate in Euro. Each participant is randomly assigned to one of the treatment combinations. Blocks with and without displayed raking were randomized within each treatment.

	PR10		PR15		PR05	
€ per item in Domain A/B	10/10		10/15		10/5	
Ranking	No	Yes	No	Yes	No	Yes
Periods	2	2	2	2	2	2
Seconds per period	120	120	120	120	120	120
Sessions	within-subjects		within-subjects		within-subjects	
No. Subjects	93		89		104	

comparison: the ranking was anonymous, private, and had no monetary consequences. For this reason, it potentially addresses self-image concerns but not reputation or status, signaling, and learning from others.[4]

As a second treatment variable, we varied monetary incentives for Domain B. The piece-rate for each correctly solved item in the prosocial task of Domain A was € 10 (donated to UNICEF) across all treatments. The corresponding piece-rate y in Domain B was € 5 in treatment PR05, € 10 in treatment PR10 (which serves as baseline), and € 15 in treatment PR15. This treatment variation was implemented between-subjects with random assignment. Thus, combining the between-subject treatments with varying piece-rates, and the within-subject treatments with implementing a ranking or not, we arrive at six treatments outlined in Table 1.

After each period subjects received feedback on the number of items solved in Domains A and B. In the corresponding ranking treatments, subjects also received information on their rank in Domain B (see Appendix C for details on the feedback screen). Each period participants started anew, i.e., items and rankings did not accumulate.

We included the following self-reported and non-incentivized questionnaire items: First, subjects had to indicate the relevance of the donation to UNICEF ("How important it is for you personally to make an effort to create donations to UNICEF for measles vaccine?") on a scale from 1 ("not important at all") to 5 ("very important") before they started solving APM items. With this question we assessed a subject's intrinsic motivation for Domain A, i.e., the prosocial activity.

[4] Since the pre-sampled professionals only solved items for Domain B, one might expect low observed rankings by design. However, the average number of correctly solved items over the two periods (in either domain) in the pre-sampled group is 4.25, which is lower than the average in any of the treatments in the main experiment (see Fig. 2 below). The median over all medians of achieved ranks out of the achievable ones (RANK $= \{1, 2, 7, 9, 12\}$) over both periods per subject is 9.5 (interquartile range: [7;12]).

After the experiment, we included the 5-item WOFO questionnaire on competitiveness (Helmreich and Spence, 1978), and the SOEP risk elicitation question on general risk taking (Dohmen et al., 2011).[5] Finally, we asked for participants' age, gender and job description.[6]

2.2. Implementation of the experiment

The online experiments (programmed in *oTree*; Chen et al., 2016) were run in two waves: In wave 1 we collected data of 12 subjects for the ranking as described above. In wave 2, 286 subjects completed the main experiment. Of our participants, 10% were female and the mean age was 37.84 years ($SD = 8.61$). We used proprietary contacts from our BEFORE database (Behavioral Finance Online Research *before.world*) to recruit finance professionals from different EU countries and across a variety of job functions. Since our study is motivated by social comparison and strong status cultures, we chose to go beyond the standard subject pool of students and focus on the finance industry with their prevalent incentive and ranking culture (Kirchler et al., 2018, 2020). Thus, we are confident that running the experiments with finance professionals increases external validity of our results and its interpretation. Given the low number of female participants in our experiment – which is typical for a finance professionals subject pool (see, e.g., Kirchler et al., 2018; Weitzel et al., 2020) – we cannot say much about gender effects (see, e.g., Murad et al., 2019). However, this is not the main question of our study. Rather, we primarily focus on professionals from an industry with a prominent ranking and competitive culture. We paid out one randomly selected period out of four. Average payment including the participation fee (€10) was €27.26 ($SD = 20.49$; $Min = 10$, $Max = 130$) for a median duration of the experiment of 15 minutes. Hence, with an average hourly wage of more than € 100, we believe that the experiment was well-incentivized.

2.3. Hypotheses

We derive hypotheses with respect to two measures: First, TOTAL denotes the total number of correctly solved items in both Domains, A and B, which we regard as a proxy for the total effort spent by an individual. Second, FRAC(B) denotes the fraction of correct answers in Domain B relative to the total number of correct answers in both Domains, A and B. This is a measure of the fraction of output achieved in Domain B and, thus, serves as a proxy for relative effort put into Domain B.[7]

Now suppose an individual faces convex costs of effort, is intrinsically motivated to spend effort in Domain A, and receives a piece-rate per unit of effort in Domain B. Then, her choice of efforts equilibrate marginal costs and benefits in the two domains. As a result, a more pronounced intrinsic motivation (proxied by the self-assessed relevance of the donation) enhances total effort as well as relative effort in Domain A (at the expense of relative effort in Domain B). In Appendix A, we specify a model of multi-task decision-making and formally derive all hypotheses.[8]

Hypothesis 1. (i) TOTAL is increasing in the self-assessed relevance of the donation; (ii) FRAC(B) is decreasing in the self-assessed relevance of the donation.

Similarly, a higher piece-rate raises total effort and relative effort spent in Domain B.[9]

Hypothesis 2. (i) TOTAL is increasing in the piece-rate; (ii) FRAC(B) is increasing in the piece-rate.

In contrast to the straightforward impact of intrinsic motivation and piece-rate, the impact of introducing a ranking depends on the desirability of a high or low rank and how this desirability is balanced or highlighted over time. If a high rank is regarded as an indication of a high overall effort rather than a poor effort in the prosocial task (in the model this is implemented by an additional utility derived from social comparison that is monotone increasing in the ranking position), the introduction of a ranking has the same impact as an enhanced piece-rate.[10]

Hypothesis 3. (i) TOTAL is increasing in the presence of a ranking; (ii) FRAC(B) is increasing in the presence of a ranking.

[5] The WOFO questions read as follows: "I enjoy working in situations involving competition with others"; "It is important to me to perform better than others on a task"; "I feel that winning is important in both work and games"; "It annoys me when other people perform better than I do"; "I try harder when I am in competition with other people". Each of these WOFO questions have been answered on a scale from 1 ("strongly disagree") to 5 ("strongly agree"). The SOEP question read as follows: "How do you see yourself: Are you generally a person who is fully prepared to take risks or do you try to avoid taking risks?" and could be answered on a scale from 0 ("risk-averse") to 10 ("risk-prone").

[6] For an overview over participants' self-reported characteristics see Table B1 in Appendix.

[7] We chose the number of correct answers instead of the number of attempts to solve items as a proxy to measure performance. The reason is that we cannot reasonably distinguish between answers provided with effort and simple tries without any effort provided ("clicking through-behavior"), which would render a variable that focuses on attempts noisy.

[8] Hypothesis 1 follows from Proposition 1 in Appendix A.

[9] For a formal derivation see Proposition 2 in Appendix A.

[10] For a formal derivation see Corollary 1 in Appendix A.

However, the individual can face a trade-off between a high rank (and correspondingly high earnings) and prosociality when deciding on total effort and in particular effort division. If the individual highlights one of the motives (implemented by a utility function that is either linear or convex in ranks), we expect the introduction of a ranking to have the same qualitative impact over periods: If a high rank is (un)desirable, introducing the ranking will have the same impact as an enhanced (reduced) piece-rate. However, if the individual balances the different motives (as modeled with a concave utility derived from ranks over time), we expect her to increase total effort and relative effort in Domain B whenever previous ranks had been sufficiently bad and to lower total effort and relative effort in Domain B if previous ranks had been sufficiently good.[11]

Hypothesis 4. (1) If $\text{RANK}_1 \ldots \text{RANK}_t$ is sufficiently bad: (i) TOTAL is increasing in the presence of a ranking; (ii) FRAC(B) is increasing in the presence of a ranking. (2) If $\text{RANK}_1 \ldots \text{RANK}_t$ is sufficiently good: (i) TOTAL is decreasing in the presence of a ranking; (ii) FRAC(B) is decreasing in the presence of a ranking.

It needs to be noted that the hypothesized impact on TOTAL vanishes whenever there is a constant rate of substitution between efforts in the two domains (e.g., if the utility from effort in the two domains is linear). In this case, changing the relevance of donation, the piece-rate, or introducing the ranking only alters the effort division but not total effort.

3. Results of the experiment

In this section we first provide descriptive results, followed by hypotheses tests and conclude with additional, exploratory analyses. Since the first three hypotheses are on aggregate behavior,[12] we pool our data for the respective analyses. This might raise concerns about order effects despite the randomization of treatments and item sequences. For this reason, we test for order effects by comparing the distribution of correctly solved items for Domain A and Domain B in treatments with ranking and without ranking, respectively, between the two treatment sequences and the two APM sequences. We conclude that there are no systematic order effects, since only one out of eight tests turns out statistically significant (for the number of correctly solved items in Domain B in treatments with ranking; Kolmogorov–Smirnov test, $N = 286$, $p < 0.05$).

Descriptive results. First, we show that the randomization of subjects into treatments resulted in the expected heterogeneity of types in each treatment: Pairwise tests of distributions (Kolmogorov–Smirnov tests) of self-reported relevance of donation (DONATION) do not reveal significant differences between piece-rate treatments ($p > 0.05$, $N \geq 182$ for each test). Distributions of normalized answers are depicted in Figure B1 in the Appendix B.

Second, we provide a first overview over the two main variables of interest, TOTAL and FRAC(B) in Fig. 2.[13] We do not find major differences in the aggregate numbers for total effort (TOTAL) across treatments. In contrast, a higher piece-rate seems to have an impact on the fraction of solved items for the selfish activity (Domain B): In Treatment PR15, FRAC(B) is higher than 0.6, exceeding the fractions in Domain B in the other treatments. The ranking provided in treatments of type RANKING shows no overall effect on TOTAL and FRAC(B).

Furthermore, we show the means of TOTAL and FRAC(B) over all subjects and piece-rates as a function of the self-assessed relevance of the donation (DONATION) in Fig. 3. For the sake of comparability across different scales, we standardize the questionnaire variables (with $ME = 0$ and $SD = 1$). In particular, we subtract the mean from each value and divide it by the standard deviation. For the questions on competitiveness (COMPETITIVENESS), we normalize each question separately before computing mean aggregated competitiveness scores which then are also normalized.[14] We find that the self-assessed relevance of the donation shows explanatory power regarding the relative effort provided in Domain B. From this figure, one might infer that a substantial fraction of subjects may be refusing to spend effort in the prosocial activity of Domain A or, put differently, invest all effort into one's own payment (Domain B). In fact, 39.86% (41.26%) of subjects spend all their effort in Domain B, i.e., FRAC(B) $= 1$, in treatments without (with) ranking. In contrast, 31.11% (30.07%) of subjects refuse to spend any effort in Domain B over both periods, i.e., FRAC(B) $= 0$, in treatments with (without) ranking (see also Figure B3 in Appendix B).

Result on Hypothesis 1. On aggregate, the total effort provided (TOTAL) is independent of the self-assessed relevance of the donation, while the relative fraction of effort put in the selfish activity in Domain B (FRAC(B)) decreases with the self-assessed relevance of the donation.

To test Hypothesis 1, we run OLS regressions with TOTAL and FRAC(B), respectively, as dependent variables.[15] We add binary treatment indicators PR15 and PR05 as explanatory variables (i.e., treatment PR10 serves as baseline), a binary variable denoting

[11] For a comprehensive discussion of modeling the utility or costs derived from a ranking and a formal statement of these observations and Hypothesis 4 see Proposition 3 in Appendix A.

[12] See also the model underlying our hypotheses, outlined in Section A in the Appendix.

[13] See Figure B2 in Appendix B for mean levels of items solved in Domain A and B, respectively. Note that the variable FRAC(B) is set to 0 in case a subject's answers are all wrong. Over the two periods this is not the case for any of our subjects, thus the results on Hypotheses 1 and 2 are not affected by this choice. However, in 28 cases subjects failed to answer any of the items correctly in one of the two periods. Reassuringly, our results on Hypotheses 3 and 4 remain robust when dropping these observations (instead of setting FRAC(B) to 0). In addition to the analysis below, please also refer to panel regression results reported in Table B3 in Appendix B.

[14] *Cronbach's Alpha* of the five normalized COMPETITIVENESS items is 0.82, indicating high internal consistency of the separate questions.

[15] Please note that the variable FRAC(B) is distributed in an interval from 0 and 1. However, for the sake of interpretability of coefficients, we report linear OLS regression results in the main text and fraction probit regression results in Table B2 in Appendix B. Results are qualitatively robust to the choice of the model.

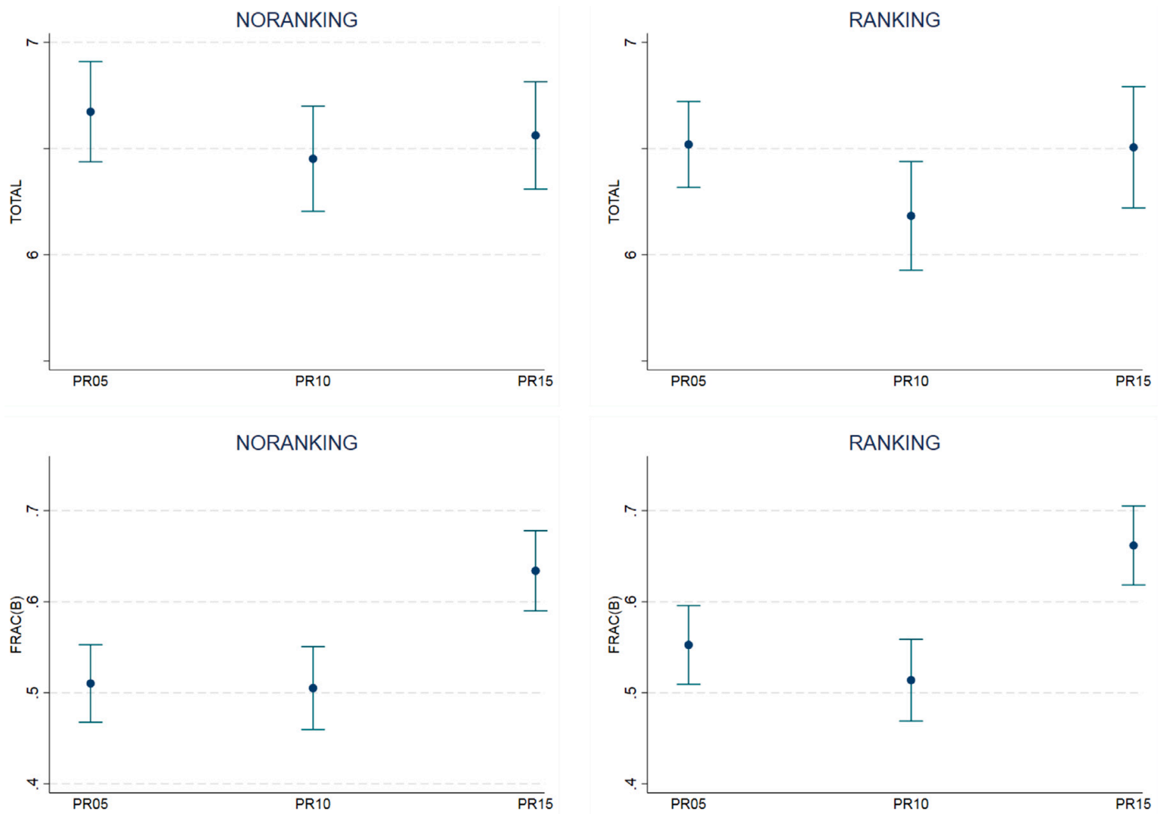

Fig. 2. Upper panel: TOTAL denotes the total number of solved items in both Domains, A and B, across the two periods. Lower panel: FRAC(B) stands for the fraction of correct items in Domain B relative to the total number of correct items in both domains. PR05, PR10, and PR15 represent the different treatments with RANKING (right panel) and without (left panel). Points indicate means, bars standard errors.

treatments with ranking (RANKING), and a variable controlling for self-reported relevance of Domain A (DONATION). We further add CONTROLS in all regressions of the paper except if otherwise noted, including gender, age, self-reported competitiveness, and self-reported risk tolerance. As can be seen from Table 2, TOTAL is independent of the self-assessed relevance of donation (DONATION), indicating no support for Hypothesis 1(i) and suggesting constant marginal returns to effort in the individual's utility function. In line with Hypothesis 1(ii), we show that the relative effort put in the selfish activity of Domain B (FRAC(B)) substantially decreases with donation relevance (DONATION).

Result on Hypothesis 2. On aggregate, neither total effort provided (TOTAL), nor the effort provided for the selfish activity in Domain B (FRAC(B)) is significantly influenced by varying piece-rates.

As outlined in Table 2, we find no evidence that the piece-rates have an impact on the total effort level (TOTAL) or on relative effort in Domain B (FRAC(B)) (see also Eckartz et al., 2012).[16] While the lack of support for Hypothesis 2(i) can be explained with constant marginal returns to effort (see the discussion at the end of Section 2.3), the failure to confirm Hypothesis 2(ii) could suggest that piece-rate variations are small relative to the individuals' intrinsic motivation to spend effort a and her concerns regarding the ranking.

With regards to Hypotheses 1 and 2, we conclude that the individual preference for the donation has an impact on relative effort provided in Domain B, whereas we find no clear evidence that the monetary incentives (piece-rates) have an effect. Our findings are particularly noteworthy because of the absence of a substitution effect of monetary incentives.[17]

Result on Hypothesis 3. The introduction of a ranking does not increase TOTAL and FRAC(B) on aggregate.

As outlined in Fig. 2 and Table 2, the introduction of a ranking does not result in an overall increase in total effort (TOTAL) or in a higher relative effort in Domain B (FRAC(B)). In other words, we do not find an aggregate effect of rank incentives. While the absence of an effect on TOTAL can again be explained by constant marginal returns to effort, the absence of an aggregate effect on FRAC(B) can be explained by a heterogeneity of ranking concerns as discussed in Section 2.3.

[16] Fig. 2 suggests that there is indeed a piece-rate effect on FRAC(B). However, this effect disappears in the analysis reported in Table 2 due to the inclusion of control variables.

[17] See also Figure B3 in Appendix B depicting the fraction of subjects refusing to spend any effort in Domain B, i.e., FRAC(B) = 0.

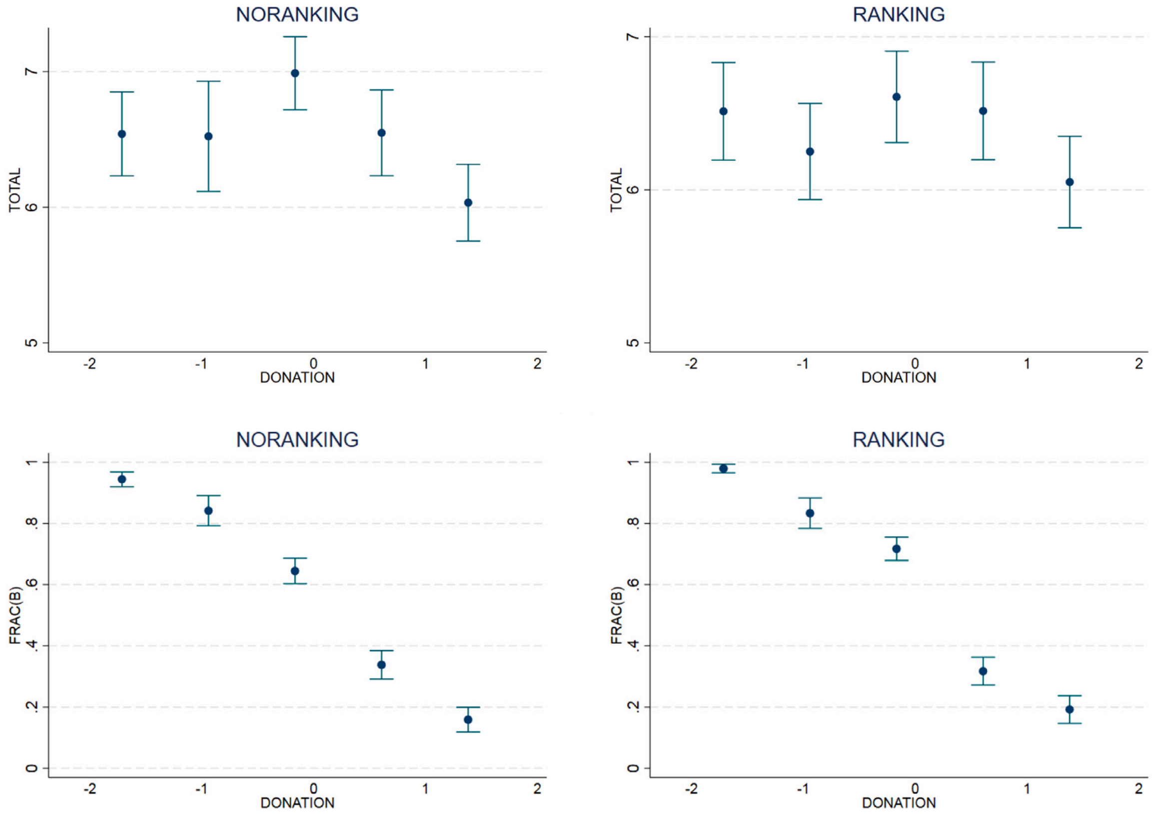

Fig. 3. Mean TOTAL, i.e. total number of solved items in both Domains, A and B, (upper panel) and mean FRAC(B), i.e., the fraction of correct items in Domain B relative to the total number of correct items in both domains, (lower panel) as a function of the self-assessed relevance of the donation (DONATION). The right panel depicts observations from treatments with RANKING and the left panel from treatments without RANKING. Points indicate means, bars standard errors.

Table 2
OLS regression results with TOTAL and FRAC(B) as dependent variables. TOTAL denotes the total number of correctly solved items in both domains, A and B, across both periods. FRAC(B) stands for the fraction of correctly solved items in Domain B relative to the total number of correctly solved items across both domains and both periods. DONATION indicates the subject's relevance of the donation (in Domain A). PR05 and PR15 are dichotomous treatment indicators. Treatment PR10 serves as baseline. The binary variable RANKING indicates the ranking treatments. Controls include gender, age, self-reported competitiveness, and self-reported risk tolerance. Standard errors, clustered on the individual level, are provided in parenthesis. * $p < 0.05$, ** $p < 0.01$.

	TOTAL	FRAC(B)
PR05	0.324	0.008
	(0.283)	(0.041)
PR15	0.145	0.078
	(0.319)	(0.045)
RANKING	−0.161	0.027
	(0.155)	(0.015)
DONATION	−0.069	−0.264**
	(0.102)	(0.015)
CONSTANT	7.718**	0.715**
	(0.593)	(0.085)
N	572	572
R^2	0.048	0.455
$P > F$	0.021	<0.001

Result on Hypothesis 4. The changes in the fraction invested in the selfish activity of Domain B from period 1 to period 2 are positively related to the position in the ranking. Those lagging behind in the ranking in period $t = 1$ put more effort in Domain B in $t = 2$. This indicates a "ranking substitution effect" of underperformers, resulting in negative consequences for the prosocial activity. The effect, however, is the opposite for those ahead in the ranking. We find no such effect for the total effort provided in both domains.

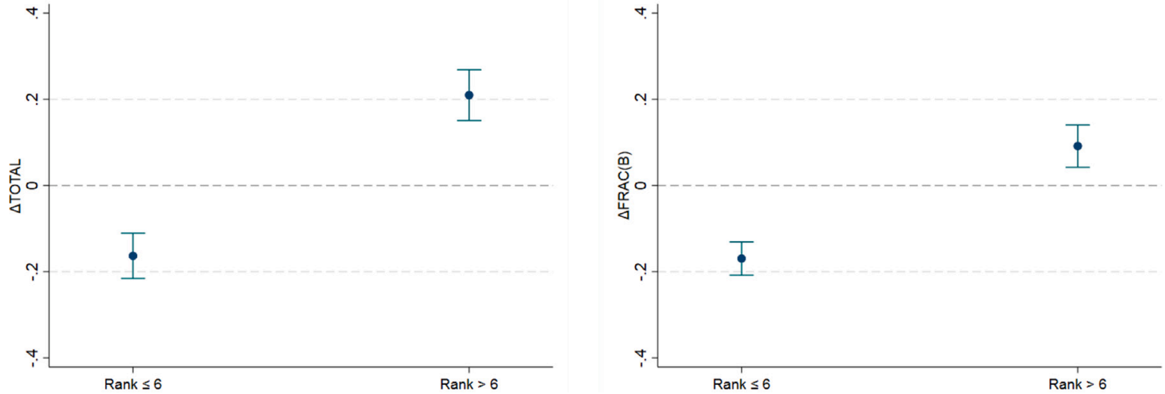

Fig. 4. Percentage changes in the number of solved items in total (ΔTOTAL; left) and in the fraction of solved items in Domain B (ΔFRAC(B); right) from period 1 to period 2 for outperformers (ranks ≤ 6) and underperformers (ranks > 6) in period 1. Points indicate means, bars standard errors.

To test Hypothesis 4 we examine percentage changes in total effort (TOTAL) and relative effort in Domain B (FRAC(B)) from period 1 to period 2 given the observed rank in the first period: $\Delta\text{TOTAL} = \frac{\text{TOTAL}_{t=2} - \text{TOTAL}_{t=1}}{\text{TOTAL}_{t=1}}$ and $\Delta\text{FRAC(B)} = \frac{\text{FRAC(B)}_{t=2} - \text{FRAC(B)}_{t=1}}{\text{FRAC(B)}_{t=1}}$.[18] Importantly, since in our experiment subjects face a random sequence of APM items, it may be that periods 1 and 2 vary in the difficulty of the items. Hence, for calculating ΔTOTAL, we adjust variable TOTAL_t by normalizing it with the average number of correct items over all participants in the same period in the same treatment for the same random sequence of APM items. ΔTOTAL measures the change in total effort from period 1 to period 2 and can be interpreted as above- and below-average changes in total effort. Since the variable FRAC(B) captures the fraction of effort spent in Domain B for each subject, no such adjustment for the difficulty of items is necessary. Thus, it is important to keep in mind that the variables TOTAL and, in turn, ΔTOTAL are calculated differently when testing Hypothesis 4 compared to the rest of the paper.

Fig. 4 provides first evidence on Hypothesis 4: We observe that those at the top of the ranking (ranks ≤ 6), on average, reduce total effort and decrease the fraction of solved items in Domain B, while those lagging behind (ranks > 6), on average, do exactly the opposite. They increase total effort and switch to the selfish activity of Domain B more frequently. While this figure provides a first visual hint on the effect of a ranking, it cannot be ruled out that the observed dynamics in ΔTOTAL can be attributed to a regression to the mean-effect or subjects' balancing between the two domains. In such a case, the observed dynamics in ΔFRAC(B) would occur without an explicitly provided ranking, for instance due to an implicit comparison to expected outcomes, ranks, or non-ranking effects such as moral licensing, regret, or feedback.[19]

Thus, we test the ranking effect conjectured in Hypothesis 4 as a treatment comparison. To do so, we run OLS regressions with ΔTOTAL and ΔFRAC(B), respectively, serving as dependent variables (see Table 3). The main explanatory variables are the binary variable denoting treatments with ranking (RANKING) and the binary variable UNDERPERFORM, which takes the value 1 in case a subject's performance is below average and zero otherwise. In treatments with ranking, a subject's performance is below average with a rank of 7 or higher; in treatments without ranking the same level of underperformance is determined by less than 4 correct items for Domain B per period. Both specifications of underperformance are equivalent, i.e., a rank of 7 or lower is achieved by solving less than 4 items for Domain B.[20] With this variable we can examine whether "outperformers" and "underperformers" react differently to the observed performance. The interaction between the treatment and underperformance variable is of particular interest, since it specifies the effect of showing a ranking *in addition* to the baseline treatment — in other words, it provides evidence for a ranking effect controlled for a potential regression to the mean-effect or balancing of motives for both domains.[21]

As shown in Table 3, we find that the period change in the absolute number of solved items not being different between treatments as the interaction term RANKING*UNDERPERFORM is small and not statistically significant. This could be an indicator of a regression to the mean or balancing effect also existent in treatments without ranking (where the underperformer dummy is indeed significant). When focusing on the relative number of solved items (ΔFRAC(B)), we find a significant treatment effect of the ranking. Those lagging behind in the ranking increase relative effort compared to those being ahead in the treatment without ranking. On average, this effect almost doubles in size in treatments with an explicit ranking, as indicated by the interaction term

[18] Please note, if $\text{TOTAL}_{t=1} = 0$ then $\Delta\text{TOTAL} = \text{TOTAL}_{t=2}$ and if $\text{FRAC(B)}_{t=1} = 0$, then $\Delta\text{FRAC(B)} = \text{FRAC(B)}_{t=2}$.

[19] For visual evidence that such dynamics are indeed relevant in treatments without ranking see Figure B5 in Appendix B.

[20] Please note that we chose the split in above- and below average performance, because subjects compete against a pre-sampled selection of peers. Therefore, for an individual subject the expected value that splits above- and below-average performance are ranks 6 and 7. Results on an specification of OLS regression models with the individual's rank in period 1 ($\text{RANK}_{t=1}$) as explanatory variable instead of UNDERPERFORM are reported in Table B5 in Appendix B. In these regression models Variable $\text{RANK}_{t=1}$ indicates a subject's rank in period 1 or, for treatments without ranking, the corresponding number of correct items ($Rank = 1$ corresponds to $B \geq 5$, $Rank = 2$ to $B = 4$, $Rank = 7$ to $B = 3$, $Rank = 9$ to $B = 2$, and $Rank = 12$ to $B <= 1$). Results remain qualitatively similar. See also o Figure B4 in Appendix B.

[21] See Table B4 in Appendix B for regression models separated for each treatment.

Table 3
OLS regression results with the following dependent variables: percentage changes in the number of solved items (ΔTOTAL) and changes in the fraction of solved items in Domain B (ΔFRAC(B)) from period 1 to period 2. The binary variable RANKING indicates the ranking treatments Variable UNDERPERFORM is a binary variable for underperforming subjects in period 1 ($Rank > 6$ or $B < 4$, respectively). PR05 and PR15 indicate treatments, i.e., PR10 serves as baseline. DONATION stands for the self-reported relevance of the donation. Controls include gender, age, self-reported competitiveness, and self-reported risk tolerance. Standard errors, clustered on the individual level, are provided in parenthesis. * $p < 0.05$, ** $p < 0.01$.

	ΔTOTAL	ΔFRAC(B)
RANKING	−0.013	−0.082
	(0.074)	(0.047)
UNDERPERFORM	0.366**	0.164**
	(0.077)	(0.059)
RANKING*UNDERPERFORM	0.053	0.143*
	(0.110)	(0.072)
PR05	0.174**	0.071
	(0.064)	(0.057)
PR15	0.057	0.059
	(0.071)	(0.066)
DONATION	−0.040	−0.032
	(0.029)	(0.022)
CONSTANT	−0.079	−0.219
	(0.135)	(0.130)
N	572	572
R^2	0.072	0.039
$P > F$	<0.001	0.007

RANKING*UNDERPERFORM (see also Table B4 in the Appendix). Thus, subjects in the ranking treatments react to the ranking relative to their peers presented in period 1: Those having a below average relative performance (i.e., a bad rank) in period 1 substitute between both domains by increasing the fraction solved in the selfish activity of Domain B at the expense of the prosocial activity — and *vice versa* for the outperforming peers.

We, thus, conclude that utility gained from social comparison is not monotone and that the reaction to observed ranks can mainly be explained by a ranking substitution effect between Domain A and Domain B. Those achieving a good rank decrease their *relative* effort for the selfish domain. In contrast, those lagging behind substitute between both domains by decreasing the number of solved items for the prosocial activity in Domain A, while increasing their relative effort in the selfish activity of Domain B. Interestingly, a similar, but significantly weaker effect also occurs without a ranking presented, which might be due to an implicit, subjective ranking or balancing motives such as moral licensing, regret, or received feedback. In contrast to substitution by domains, we do not find an effect of the observed ranking on the change in absolute number of solved items (ΔTOTAL).

Explanation of achieved ranks. Additionally and in more exploratory terms, we examine which characteristics explain a subject's rank. Since the number of correct items in Domain B directly translates into ranks, we provide OLS regression models with the number of correct items in Domain B in the main text and ordered logistic regression models using a subject's rank – providing qualitatively similar results – in Table B6 in Appendix B. As explanatory variables we include a subject's self-assessed answers regarding donation relevance, competitiveness, and risk tolerance (DONATION, COMPETITIVENESS, and RISK, respectively), as well as a binary indicator for gender (FEMALE) and the age (AGE). For the regression on the individuals' rank in period 2 in model 2 we also include the number of correct items in Domain B in period 1.

Results are reported in Table 4: To begin with, we find that the number of correctly solved items in Domain B in period 1 explains a subject's performance in Domain B in period 2. Together with the results described above, this indicates that, on average, subjects ahead in the ranking seem to manage to keep their good relative position despite the reduced effort in Domain B. Moreover, we report that in the first period subjects' self-assessed competitiveness (COMPETITIVENESS) explains their performance in Domain B, with more competitive subjects achieving a better performance. This relation becomes insignificant once the ranking is announced in period 2. Potentially, this is an indication of "crowding out" of intrinsic levels of competitiveness by the externally provided ranking. Finally, we find that higher self-reported donation relevance (DONATION) leads to a worse performance in Domain B, as more effort is invested in the prosocial activity of Domain A. The effect size is considerable, even though it substantially decreases in the second period (while remaining significant), which, again, might be due to a crowding-out effect.

4. Discussion

Our results have several policy implications that we elaborate on in this section. However, before we proceed, we would like to discuss generalizability and limitations in order to provide a critical framework for interpreting policy implications.

Table 4
OLS regression results with a subject's correct number of items in Domain B in period n ($B_{t=n}$), n indicating period 1 or 2) as dependent variable. $B_{t=1}$ is included as explanatory variable in the specification for period 2 indicating a subject's correct number of items in period 1. DONATION stands for self-reported relevance of the donation, COMPETITIVENESS for self-reported competitiveness, and RISK for self-reported risk tolerance. Robust standard errors are provided in parenthesis. * $p < 0.05$, ** $p < 0.01$.

	$B_{t=1}$	$B_{t=2}$
$B_{t=1}$		0.336**
		(0.061)
COMPETITIVENESS	0.287*	−0.053
	(0.121)	(0.090)
DONATION	−1.022**	−0.345**
	(0.094)	(0.102)
RISK	0.124	0.021
	(0.108)	(0.078)
FEMALE	0.001	0.294
	(0.344)	(0.294)
AGE	−0.026*	−0.009
	(0.011)	(0.008)
CONSTANT	3.215**	1.078
	(0.465)	(0.347)
N	286	286
R^2	0.294	0.358
$P > F$	<0.001	<0.001

4.1. Generalizability and limitations

There are reasons to expect our results to be relevant outside the lab: First, since we find ranking effects in anonymous online experiments, we might expect even more pronounced effects in real life settings where (public) status plays an important role and can affect monetary incentives and career outcomes. In addition, in our experiment it is public knowledge that every participant is a finance professional. Given that social comparison is stronger among peers, we expect our experimental design to allow for observing results with comparatively high external validity. Second, in the business world, there are plenty of situations in which individuals are already at the maximum of expanding effort and mainly focus on effort division. A working day is limited with individuals deciding on which activities to focus on. In our experiment, subjects also have a limited time budget of two minutes to solve as many items as possible and thereby we believe to mimic a crucial real-world feature with our model.

However, we also see limitations of our experimental results: First, while mirroring relevant aspects of business decisions, our experimental design clearly is an abstraction from real-life decision situations. For instance, the implementation of the ranking, in particular the pre-sampled reference group, is due to pragmatic reasons and a cleaner identification of relevant dynamics. At the same time, the ranking in our experiment differs from rankings as implemented in companies or economic sectors. On the other hand, since our focus is on the impact of social comparison, it seems realistic to assume that more egoistic people or individuals facing stronger monetary incentives tend to perform better in the incentivized dimension and, consequently, are the comparison group also in companies or within a profession. As another example of design limitation, while we consider it one of the advantages of the APM task that participants are motivated to obtain a higher rank in an intelligence test (Falk and Szech, 2019), such intrinsic motivation can also counteract extrinsic piece-rates motivation.

Second we would also like to point out limitations in our analysis, in particular regarding statistical power: Given that we chose a finance professionals subject-pool to increase external validity of our results, we are limited in the number of participants. As observations are one of the factors influencing statistical power, there is a possibility of low power in our analysis.[22] In our case, this specifically calls for caution in the interpretation of results on Hypothesis 4, since potentially low power might be aggravated when it comes to results reported in Table 3 based on an interaction effect. While we indeed find a significant effect, we might face an issue of inflated effect size. Consequently, the expected economic dimension of policy implications discussed below should be cautiously considered with this in mind. Since we deem the results and the potential policy implications interesting and relevant, we believe that further research is warranted and needed. In particular, we would consider it worthwhile to replicate our experiment to increase external validity, potentially with a different subject pool and a much higher sample size. The former could generalize the results to other relevant groups, while the latter could address power concerns.

4.2. Policy implications

Social comparison based on rankings features most prominently in labor markets with strong competition for talent (e.g., finance professionals, science, top managers etc.). One objective of rankings is information provision as it allows employers and/or peers

[22] Unfortunately, we have not conducted a power analysis before running the study.

to identify top-performers. Moreover, rankings are also often used as an incentive device: Payments are explicitly or implicitly contingent on the rank and the goal to achieve a high rank may also be a motive for better performance *per se*. Our results, however, indicate hidden costs of the introduction of rankings and potential lessons to be learned for the design of incentive schemes.

Hidden costs of rankings. In contrast to previous studies on the impact of piece-rates and rankings in single-task environments, our investigation of a multi-tasking environment suggests two yet unrecognized effects:

Ranking Substitution Effect: Neither an exogenous variation of piece-rates nor the considerable heterogeneity of individual's self-reported assessment of the relevance of the donation nor the introduction of a ranking have a significant impact on *total* effort in our experiment. However, participants *substitute* efforts spent in the two domains. The lower the relevance of the donation and the higher the piece-rate, the higher is the fraction of effort spent in the domain that determines the agent's earnings (Domain B). While this does not imply that more substantial piece-rate variations may not also have an impact on total effort, it clearly indicates that any attempt to increase total effort with higher piece-rates in Domain B would at least come at the cost of significant substitution of efforts across domains.

In contrast, a ranking neither influences total nor relative effort on the aggregate level in our experiment. This is because the impact of a ranking on an individual's decision crucially depends on the individual's rank. Individuals lagging behind in the ranking tend to substitute effort in the prosocial task with effort in the task that determines earnings and the ranking position. For these individuals, a ranking has the same effect as a higher piece-rate. In contrast, individuals leading in the ranking substitute effort in the ranked activity with prosocial effort, i.e., these individuals react as if the donation has become more relevant.

Ranking Attenuation Effect: If – in response to a ranking – an individual with a bad rank spends more effort in Domain B while an individual with a good rank spends less effort in Domain B, the difference in effort division across individuals tends to diminish because of the ranking. This is in stark contrast to the impact of rankings in single-task environments (without a prosocial dimension) where a ranking and tournament incentives are often used with the intention to increase effort differentials between agents who differ in productivity.[23] In this case, more productive agents (i.e., agents with lower effort costs or higher returns to a given piece-rate) spend *more* effort compared to less productive agents in the presence of a ranking which yields a larger difference in total and relative effort in response to a ranking.

Implications for contract design. Potential substitution and attenuation effects have several implications for the optimal design of incentive schemes.

(1) Agency Costs: For single-task environments it is often assumed that a given effort by agents can be achieved with a lower piece-rate (i.e., at lower costs) in the presence of a ranking. Our findings for a multi-tasking environment with a prosocial dimension might indicate that this is only true for agents lagging behind in the ranking. For outperforming agents, the ranking rather seems to operate similar to a piece-rate reduction. Hence, lowering the piece-rate and introducing a ranking may lead to the same total and relative effort provision by underperforming agents, but might introduce a disincentive for relative effort for agents leading in the ranking. As a consequence, the introduction of a ranking does not unambiguously reduce agency costs — it might work as a rather imperfect substitute for piece-rates.

(2) Bonus Caps: Receiving similar efforts for reduced piece-rates in the presence of a ranking is not only reducing agency costs, it may also be regarded as a valuable tool for maintaining incentives in the presence of payment regulations. Consider, for example, the frequently discussed cap of bonus payments (see, e.g., Bénabou and Tirole, 2016): If piece-rates are regulated to be below a certain threshold value, a ranking would be a tool to further increase total and relative effort into the task that is subject to the ranking. Our findings suggest that offsetting (regulated) piece-rates by a ranking may induce significant substitution across tasks and disincentives for outperforming agents. If these substitution and attenuation effects are sufficiently pronounced, introducing the ranking may be inferior not only from a welfare perspective, but also from the firm's point of view.

(3) Selecting productive agents: In a single-task environment, it is often assumed that introducing a ranking enhances total efforts and increases the difference between efforts spent by more or less productive and/or intrinsically motivated agents for a given piece-rate. As a consequence, offering contract menus tailored to more or less productive agents becomes more attractive in the presence of a ranking. In contrast, potential substitution and attenuation in a multi-tasking environment can reduce the effect of a ranking, potentially also on the difference between total and relative efforts by agents who lead or lag behind in the ranking. This makes the selection of agents with different productivity and/or intrinsic motivation not only less attractive but also less feasible. Ulrichshofer and Walzl (2020) add utility from social comparison as in the model described in Appendix A to the set-up by Bénabou and Tirole (2016) and analyze how substitution and attenuation alters optimal screening contracts depending on the intensity of labor market competition. If employers have high market power, a psychological cost from lagging behind in a ranking can reduce efficiency distortions of optimal screening contracts, as contracts with a low piece-rate become less attractive for high productivity agents. If, however, the competition intensity is high, psychological costs of lagging behind only amplify efficiency distortions, as contracts with a high piece-rate become more attractive for low productivity agents. In contrast – and for the same reason – psychological costs of leading in the ranking can enhance efficiency in this case.

[23] As noted in Section 1, however, there is evidence of effects of rankings on effort, depending, among others, on expectations, current rank, and the principal agent relationship. See, e.g., Al-Ubaydli and List (2015), Blanes-i-Vidal and Nossol (2011) and Tran and Zeckhauser (2012).

5. Conclusion

In this paper, we examined how monetary and rank incentives influence effort in a multi-tasking environment with a trade-off decision between a monetarily incentivized and a prosocial task. Depending on how a decision-maker interprets a provided ranking on the selfish task, it can serve as an additional incentive or disincentive. If a good rank is desirable, it increases total effort and the fraction of effort spent for the ranked task. If a good rank is costly because it is regarded as a signal (to the self or others) of low prosocial activity, it lowers total effort and the fraction of effort spent in the ranked domain. If the agent balances the conflicting motives over time, we expect a ranking to be a disincentive for the ranked task if the agent's performance was ranked highly in previous periods and an incentive for the ranked task in case of a poor ranking in previous periods.

We tested these hypotheses in a controlled online experiment with 286 internationally operating finance professionals. We found that the introduction of a ranking on the monetarily incentivized activity leads to a "ranking substitution effect": Outperforming professionals substituted relative effort spent for their own payment by putting more effort in the prosocial activity. In contrast, underperforming professionals substituted by spending more effort for their own payment and less for the prosocial activity.

Finally, given that our findings indicate hidden costs of the introduction of rankings and potential lessons to be learned for the design of incentive schemes, further research on the interplay of monetary incentives, social comparison, and prosocial behavior in multi-tasking problems is needed.

Data availability

Our experimental software, data files, and the analysis script file are stored in the following repository: https://osf.io/ujncd/.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.euroecorev.2023.104458.

References

Al-Ubaydli, Omar, List, John A., 2015. On the generalizability of experimental results in economics. In: Frechette, G., Schotter, A. (Eds.), Handbook of Experimental Economic Methodology. Oxford University Press, pp. 420–463.

Anderson, Melissa, Ronning, Emily, Vries, Raymond De, Martinson, Brian, 2007. The perverse effects of competition on scientists' work and relationships. Sci. Eng Ethics 13 (4), 437–461.

Ariely, Dan, Bracha, Anat, Meier, Stephan, 2009. Doing good or doing well? Image motivation and monetary incentives in behaving prosocially. Amer. Econ. Rev. 99, 544–555.

Azmat, Ghazala, Iriberri, Nagore, 2010. The importance of relative performance feedback information: Evidence from a natural experiment using high school students. J. Publ. Econ. 94 (7–8), 435–452.

Baker, George, 1992. Incentive contracts and performance measurement. J. Polit. Econ. 100 (3), 598–614.

Ball, Sheryl, Eckel, Catherine, Grossman, Philip, Zame, William, 2001. Status in markets. Q. J. Econ. 116 (1), 161–188.

Bandiera, Oriana, Barankay, Iwan, Rasul, Imran, 2010. Social incentives in the workplace. Rev. Econom. Stud. 77, 417–458.

Barankay, Iwan, 2015. Rank incentives: Evidence from a randomized workplace experiment. Working Paper.

Bebchuk, L., Spamann, H., 2010. Regulating bankers' pay. Georgetown Law J. 98 (2), 247–287.

Bénabou, Roland, Tirole, Jean, 2006. Incentives and prosocial behavior. Amer. Econ. Rev. 96 (5), 1652–1678.

Bénabou, Roland, Tirole, Jean, 2016. Bonus culture: competitive pay, screening, and multitasking. J. Polit. Econ. 124 (2), 305–370.

Blanes-i-Vidal, Jordi, Nossol, Mareike, 2011. Tournaments without prizes: Evidence from personnel records. Manage. Sci. 57 (10), 1721–1736.

Borghans, Lex, Meijers, Huub, ter Weel, Bas, 2013. The importance of intrinsic and extrinsic motivation for measuring IQ. Econ. Educ. Rev. 34, 17–28.

Bowles, Samuel, Polania-Reyes, Sandra, 2012. Economic incentives and social preferences: Substitutes or complements? J. Econ. Lit. 50 (2), 368–425.

Brown, Keith C., Harlow, W.V., Starks, Laura T., 1996. Of tournaments and temptations: An analysis of managerial incentives in the mutual fund industry. J. Finance 51 (1), 85–110.

Charness, Gary, Masclet, David, Villeval, Marie Claire, 2014. The dark side of competition for status. Manage. Sci. 60 (1), 38–55.

Chen, D.L., Schonger, M., Wickens, C., 2016. oTree—An open-source platform for laboratory, online, and field experiments. J. Behav. Exper. Finance 9, 88–97.

Cohn, Alain, Fehr, Ernst, Götte, Lorenz, 2015. Fair wages and effort provision: combining evidence from a choice experiment and a field experiment. Manage. Sci. 61, 1777–1794.

Coles, Jeffrey L., Li, Zhichuan (Frank), Wang, Albert Y., 2018. Industry tournament incentives. Rev. Financ. Stud. 31 (4), 1418–1459.

Diamond, Douglas, Rajan, Raghuram, 2009. The credit crisis: Conjectures about causes and remedies. Amer. Econ. Rev. 99 (2), 606–610.

Dijk, Oege, Holmen, Martin, Kirchler, Michael, 2014. Rank matters – The impact of social competition on portfolio choice. Eur. Econ. Rev. 66, 97–110.

Dohmen, Thomas, Falk, Armin, Huffman, David, Sunde, Uwe, Schupp, Jürgen, Wagner, Gert, 2011. Individual risk attitudes: Measurement, determinants, and behavioral consequences. J. Eur. Econom. Assoc. 9 (3), 522–550.

Eckartz, Katharina, Kirchkamp, Oliver, Schunk, Daniel, 2012. How do Incentives affect Creativity?. Working Paper.

Falk, Armin, Szech, Nora, 2019. Competing Image Concerns: Pleasures of Skill and Moral Values. Working Paper.

Fanelli, Daniele, 2010. Do pressures to publish increase scientists' bias? An empirical support from US states data. PLoS One 5, 1–7.

Festinger, Leon, 1954. A theory of social comparison processes. Hum. Relations 7, 117–140.

Fishbach, Ayelet, Zhang, Ying, Koo, Mingjung, 2009. The dynamics of self-regulation. Eur. Rev. Soc. Psychol. 20, 315–344.

Gill, David, Kissova, Zdenka, Lee, Jaesun, Prowse, Victoria, 2019. First-place loving and last-place loathing: How rank in the distribution of performance affects effort provision. Manage. Sci. 65 (2), 494–507.

Gneezy, Uri, List, John, Livingston, Jeffrey, Qin, Xiangdong, Sadoff, Sally, Xu, Yang, 2019. Measuring success in education: The role of effort on the test itself. Am. Econ. Rev. Insights 1 (3), 291–308.

Gneezy, Uri, Rustichini, Aldo, 2000a. A fine is a price. J. Legal Stud. 29 (1), 1–17.

Gneezy, Uri, Rustichini, Aldo, 2000b. Pay enough or don't pay at all. Q. J. Econ. 115 (3), 791–810.

Harrison, Glenn, List, John, 2004. Field experiments. J. Econ. Lit. 42, 1009–1055.

Helmreich, Robert L., Spence, Janet T., 1978. The work and family orientation questionnaire: An objective instrument to assess components of achievement motivation and attitudes toward family and career. Catalog Sel. Documents Psychol. 8 (2), Document 1677.
Kaniel, Ron, Parham, Robert, 2017. WSJ category kings – The impact of media attention on consumer and mutual fund investment decisions. J. Financ. Econ. 123 (2), 337–356.
Kirchler, Michael, Huber, Jürgen, Stefan, Matthias, Sutter, Matthias, 2016. Market design and moral behavior. Manage. Sci. 62 (9), 2615–2625.
Kirchler, Michael, Lindner, Florian, Weitzel, Utz, 2018. Rankings and risk-taking in the finance industry. J. Finance 73 (5), 2271–2302.
Kirchler, Michael, Lindner, Florian, Weitzel, Utz, 2020. Delegated investment decisions and rankings. J. Bank. Financ. 120.
Kuhnen, Camelia, Tymula, Agnieszka, 2012. Feedback, self-esteem, and performance in organizations. Manage. Sci. 58 (1), 94–113.
Kuziemko, Ilyana, Buell, Ryan, Reich, Taly, Norton, Michael I., 2014. "Last-place aversion": Evidence and redistributive implications. Q. J. Econ. 129 (1), 105–149.
Lozano, Jose H., Revuelta, Javier, 2020. Investigating operation-specific learning effects in the Raven's advanced progressive matrices: A linear logistic test modeling approach. Intelligence 82, 101468.
Mas, Alexandre, Moretti, Enrico, 2009. Peers at work. Amer. Econ. Rev. 99 (1), 112–145.
Mellström, Carl, Johannesson, Magnus, 2008. Crowding out in blood donation: Was titmuss right? J. Eur. Econom. Assoc. 6 (4), 845–863.
Moldovanu, Benny, Sela, Aner, Shi, Xianwen, 2007. Contests for status. J. Polit. Econ. 115 (2), 338–363.
Murad, Zahra, Stavropoulou, Charitini, Cookson, Graham, 2019. Incentives and gender in a multi-task setting: An experimental study with real-effort tasks. PLoS One 14 (3), 1–18.
Osterloh, Margit, Frey, Bruno, 2015. Ranking games. Evaluat. Rev. 39 (1), 102–129.
Rajan, Raghuram G., 2006. Has finance made the world riskier? Eur. Financial Manag. 12 (4), 499–533.
Raven, John, 2000. The Raven's progressive matrices: Change and stability over culture and time. Cogn. Psychol. 41 (1), 1–48.
Shleifer, Andrei, 2004. Does competition destroy ethical behavior? Am. Econ. Rev. 94, 414–418.
Sirri, Erik R., Tufano, Peter, 1998. Costly search and mutual fund flows. J. Finance 53 (5), 1589–1622.
Tran, Anh, Zeckhauser, Richard, 2012. Rank as an inherent incentive: Evidence from a field experiment. J. Publ. Econ. 96, 645–650.
Ulrichshofer, Anna, Walzl, Markus, 2020. Social Comparison and Optimal Contracts in the Competition for Managerial Talent.
Weitzel, Utz, Huber, Christoph, Huber, Jürgen, Kirchler, Michael, Lindner, Florian, Rose, Julia, 2020. Bubbles and financial professionals. Rev. Financ. Stud. 33 (6), 2659–2696.