Senn, Julien; Schmitz, Jan; Zehnder, Christian

Working Paper

Leveraging social comparisons: The role of peer assignment policies

Working Paper, No. 427

Provided in Cooperation with:
Department of Economics, University of Zurich

Suggested Citation: Senn, Julien; Schmitz, Jan; Zehnder, Christian (2023) : Leveraging social comparisons: The role of peer assignment policies, Working Paper, No. 427, University of Zurich, Department of Economics, Zurich, https://doi.org/10.5167/uzh-231484

This Version is available at:
https://hdl.handle.net/10419/275665

University of Zurich[UZH]

University of Zurich

Department of Economics

Working Paper Series

ISSN 1664-7041 (print)
ISSN 1664-705X (online)

Working Paper No. 427

Leveraging Social Comparisons: The Role of Peer Assignment Policies

Julien Senn, Jan Schmitz and Christian Zehnder

Revised version, August 2023

Leveraging Social Comparisons: The Role of Peer Assignment Policies*

Julien Senn Jan Schmitz Christian Zehnder

August 11, 2023

Abstract

Using a large-scale real effort experiment, we explore whether and how different peer assignment mechanisms affect worker performance and stress. Letting individuals choose whom to compare to increases productivity to the same extent as a targeted exogenous matching policy designed to maximize motivational spillovers. These effects are significantly larger than those obtained through random assignment and their magnitude is comparable to the impact of monetary incentives that increase pay by about 10 percent. A downside of targeted peer assignment is that, unlike endogenous peer selection, it leads to a large increase in stress. Using a combination of choice data, text analysis and simulations, we show that the key advantage of letting workers choose whom to compare to is that it allows those workers who want to be motivated to compare to a motivating peer while also permitting those for whom social comparisons have little benefits or are too stressful to avoid them. Finally, we provide evidence that social comparisons yield stronger motivational effects than comparable non-social goals.

Key Words: Social comparisons, Productivity, Stress, Incentives, Real effort
JEL Codes: C93, J24, M54

*We thank Yan Chen, Luke Coffman, Stefano DellaVigna, Benjamin Enke, Christine Exley, Armin Falk, David Huffman, Judd Kessler, Sebastian Kube, Michel Maréchal, Stephan Meier, Jonas Radbruch, Florian Schneider, Matthias Sutter, Bertil Tungodden, Florian Zimmermann, Utz Weitzel, Ulf Zölitz and numerous seminar participants for helpful comments and suggestions. Financing from the University of Zürich as well as the Swiss National Science Foundation (Grants 100018_200942 and 407340_172397) is gratefully acknowledged. The study was pre-registered in the AEA RCT Registry (AEARCTR-0003217, https://www.socialscienceregistry.org/trials/3217). Schmitz: Radboud University Nijmegen (jan.schmitz@ru.nl). Senn: University of Zurich (julien.senn@econ.uzh.ch). Zehnder: University of Lausanne (christian.zehnder@unil.ch)

1 Introduction

Social comparisons play an important role at the workplace. Workers often compare to their peers, and these comparisons tend to increase productivity (see e.g. Falk and Ichino, 2006). While these social comparisons typically arise spontaneously between coworkers, an increasing number of firms purposefully try to make them more salient—e.g. through dynamic, computerized, leaderboards.[1] But can social comparisons really be leveraged to boost productivity?

To date, the literature has almost exclusively focused on the effects of social comparisons with randomly assigned peers. In this context, motivational spillovers have been shown to often depend on the characteristics of both the observed peer and the observer (see e.g. Villeval, 2020), suggesting that it might indeed be possible to further leverage social comparisons. In particular, alternative peer assignment mechanisms that systematically expose workers to particularly motivating peers might boost performance beyond what can be achieved by randomly assigned peers.

A simple alternative to random-assignment is to let workers choose whom they would like to compare to. As a matter of fact, such self-chosen comparisons are pervasive in many contexts (see e.g. Fujita and Diener, 1997; Suls et al., 2002), including at the workplace.[2] If workers know what type of comparison motivates them most, endogenous peer choice might be highly effective, and workers might be particularly motivated because they get to observe the peer they explicitly chose (choice effect).[3] However, the risk of such a self-governed system is that some workers might shy away from comparing to others (e.g. to avoid being distracted or stressed out) or might even choose to compare to peers that are not motivating (e.g. by comparing downwards to feel good about themselves).

Another potentially promising way to improve the productivity of the workforce in the presence of heterogeneous peer effects is to exogenously assign workers to peers that are predicted to be highly motivating, as recently theorized (see e.g. Graham et al., 2014; Roels and Su, 2014; Kräkel, 2016). Practical attempts that go in this direction can be seen in the recent trend to"gamify" the provision of information

[1] Leaderboards are (computerized) ranking systems that provide employees with information about how they compare to selected colleagues in terms of workplace performance. Such leaderboards are highly flexible: they can easily be adapted to almost any context and they can provide individualized, tailored information to different workers—also in real-time.

[2] For example, researchers compare their research output to a selected subset of their colleagues, wealth managers compare the returns of their portfolio with the returns of some of their competitors, and schoolchildren compare their grades with only a few of their classmates.

[3] Such a preference for 'chosen alternatives' over similar 'assigned alternatives' has been discussed in other other contexts (see, e.g., Dal Bó et al., 2010).

about coworkers' productivity to boost output.[4] However, such systematic, exogenous assignment procedures might backfire if workers get upset about being pressured to observe peers they would *not* have chosen (mismatch effect). Moreover, the implementation of such systems requires detailed information and can be associated with considerable costs.

Despite the relatively widespread prevalence of such non-random assignment policies in practice, very little is known about their impact on performance and their potential unintended implications. Our paper makes a first step towards filling this important gap in the literature by providing new causal evidence on the behavioral effects of social comparisons when peers are either i) randomly assigned, ii) exogenously assigned to maximize expected productivity, or iii) endogenously chosen by the workers. Moreover, while the existing literature has predominantly focused on the role of social comparisons for productivity, we also shed light on their effects on workers' perceived stress since recent evidence from psychology suggests that social comparisons might also affect psychological well-being (see e.g. Buunk and Dijkstra, 2017; Bárcena-Martín et al., 2017; Fujita and Diener, 1997).[5]

In our setting, letting individuals choose whom to compare to increases productivity to the same extent as a targeted exogenous matching policy aimed at maximizing motivational spillovers. These effects are significantly larger than those obtained through random assignment and their magnitude is comparable to the impact of introducing monetary incentives. However, whereas targeted peer assignment leads to a strong increase in perceived stress, endogenous peer choice only has a moderate impact on stress.

We uncover the behavioral origins of these desirable effects of peer choice. We show that exogenously imposed social comparisons entail a potentially important trade-off: inducing a worker to observe a peer not only creates motivational spillovers, but also increases the stress level of the observer. For virtually all categories of workers, the magnitude of both effects increases in the productivity of the observed peer, i.e. if the observed peer's productivity is higher, the motivational effect is larger, but so is the increase in perceived stress. Thus, a targeted exogenous matching policy which systematically assigns workers to the predicted most motivating

[4]This practice is being widely adopted by a large number of firms accross different industries, from sales to (online) retail and banking, among others. Examples of such companies include Target, Amazon, and Disney, to name a few (see, e.g., Financial Times: https://tinyurl.com/ycystmez). For a more general discussion of the rise of gamification and for additional concrete examples, see Koivisto and Hamari (2019).

[5]Stress is associated with a number of (negative) consequences for workers. For example, higher stress at the workplace has been associated with lower productivity (Halkos and Bousinakis, 2010), higher absenteeism (Jacobson et al., 1996; Leontaridi and Ward-Warmedinger, 2002), and higher turnover intentions (Mosadeghrad, 2013), among others.

peer generates high productivity gains, but also leads to a strong increase in stress. In addition, we also document that the motivating effect of exogenously assigned peers is substantially reduced if workers are matched with someone they would not have chosen themselves–a phenomenon we refer to as the "*mismatch effect*".

Endogenous peer choice elegantly circumvents these issues. Combining choice data with text-analysis of reported motives, we demonstrate that a large proportion of workers aims at increasing their productivity. These workers typically choose to compare to a highly productive peer and thus benefit from strong motivational spillovers, just like the workers who are exogenously assigned to the highly productive peer. At the same time, a smaller yet non-negligible proportion of workers refrains from comparing with a peer because they worry about being distracted, believe that observing a peer is irrelevant, or are concerned that such a comparison will be stressful. The decision *not* to compare themselves with a peer allows them to avoid a high increase in stress. Of course, these workers do not benefit from motivational spillovers, but it is important to keep in mind that social comparisons are not particularly motivating for people who do not wish to compare (mismatch effect). In a nutshell, the key advantage of letting workers choose whom to compare to is that it allows those workers who want to be motivated to compare to a motivating peer, but it also prevents those for whom social comparisons have little benefits from being exposed to high levels of stress.

Our work relies on a large-scale (N=6532), pre-registered real-effort experiment conducted in an online labor market. In our experiment, participants are hired to perform a simple real effort task in two consecutive periods. The first period is identical across all treatments: workers are required to complete the task in isolation. This allows us to obtain a clean measurement of each worker's baseline productivity. Each worker is then provided with (private) information about their performance relative to the performance of "*60 other participants who already took part in this study*" (the "reference population"). This relative performance feedback allows each worker to assess how good they are at the task in comparison to similar others. Workers are then randomized into different conditions at the beginning of the second period. Our treatments exogenously vary two dimensions: i) whether and how participants are matched with a peer (a "reference worker") while they are completing the task in period 2 (no reference worker, random assignment to a reference worker, targeted assignment to the predicted most motivating reference worker, or endogenous choice of a reference worker), and ii) the compensation scheme used to pay participants in period 2 (fixed wage vs. performance pay).

Our setting resembles many work environments that allow workers to observe

each other and in which such comparisons are hard to avoid. While social comparisons can, in principle, affect individuals through multiple channels, we focus on the *motivational spillovers* that arise from *observing* a peer. We therefore consciously restrict our attention to a setting that neither involves production complementarities between workers, nor provides scope for social learning.

A well-known challenge that arises when studying social spillovers in observational studies is the reflection problem (see e.g. Manski, 1993). Suppose that two workers, i and j, can observe each other while independently working on a task that involves no production complementarities. For worker i, there might be some motivational spillover (either positive or negative) from observing worker j, but worker j might also alter its productivity as a response to being observed by worker i, thereby making the identification of the motivational spillovers difficult. Our design circumvents this problem by providing workers with *real-time* information about the productivity of a reference worker that is drawn from the reference population (and whose performance can thus no longer change).

Another hurdle in the identification of social spillovers is that social comparisons also always convey information about relative performance (which we refer to as the "feedback problem"). Because performance feedback has been shown to affect productivity even in situations where no monetary incentives are at stake (Charness et al., 2013), not controlling for such feedback effects might introduce an important confound. We account for this problem by providing all our workers with information about their rank within the reference population after the first period, so that the relative performance feedback is held constant across conditions. This allows us to cleanly isolate motivational spillovers and distinguish them from informational effects.

Our analysis relies on data from seven different conditions. As the baseline condition for evaluating the causal (and possibly heterogeneous) effects of social comparisons, we use a treatment in which participants are randomly assigned to a reference worker in period 2 (EXRA). The reference workers are drawn from the reference population. To limit the number of possible comparisons, we restrict the set of potential reference workers to three possibilities: a highly productive worker, an average-productivity worker, and low productivity worker. Participants who are assigned to one of these reference workers receive *real-time* information about the performance of this reference worker while completing the task in the second round. This treatment allows us to identify the causal effects of observing a randomly assigned peer and to uncover potential heterogeneities in motivational spillovers across both observing and observed workers.

We compare this random-assignment condition to a treatment in which participants *only* get relative performance feedback (RANK). We show that exogenously assigned comparisons boost productivity beyond the effects of simple performance feedback. In addition, our data reveal that social comparisons also substantially affect participants' stress: participants exposed to a more productive reference worker not only become more productive, but they also report a much stronger increase in stress. These findings highlight that social comparisons not only have motivational potential, but may also entail non-negligible costs for the workers.

To explore the extent to which alternative assignment policies can further leverage motivational spillovers, we compare these results with two non-random matching procedures that can be implemented by practitioners: a policy in which workers are given the opportunity to choose whom to compare to (ENDO) and a targeted exogenous assignment policy in which workers are matched with their predicted most motivating peer (EXBE). These two treatments enable us to demonstrate that ENDO and EXBE similarly outperform EXRA in the productivity dimension, but that targeted matching yields substantially larger levels of perceived stress.

What explains that the productivity gains in ENDO and EXBE are virtually indistinguishable, but the impact on perceived stress is different? To answer this question several mechanisms need to be disentangled. We first explore workers' preferences for peers. We show that the set of chosen reference workers in ENDO is very different from the set of predicted most motivating reference workers in EXBE. Whereas almost all workers in EXBE are assigned to the high-performance reference worker, only about 45% of workers in ENDO choose the most productive reference worker. The second most frequent decision is not to compare to anyone (30%). Text analysis of reported choice motives unveils that this choice pattern is logically associated with different choice motives: workers who aim at improving their productivity choose to compare to the highly productive peer, whereas those who wish to avoid stress or distraction choose *not* to compare to anyone. It is therefore no surprise that ENDO generates less stress, since fewer workers compare to the most stressful peer. However, these results raise an interesting question: Given that fewer individuals compare to the predicted most motivating peer, why is the productivity in ENDO not substantially lower than in EXBE?

We explore this question using simulations that combine performance data and preferences for peers from EXRA with peer selection data from ENDO. These simulations allow us to show that the motivational potential of exogenously assigned peers is fully unleashed only if the exogenously assigned peer coincides with the worker's preferred choice. If the assigned peer does not correspond to the preferred choice

(i.e. if there is a mismatch between a workers' preference for a peer and the assigned peer), the motivating effect is reduced. This mismatch effect provides an explanation for why performance remains high in ENDO: those workers who choose not to compare to the high-performance reference worker would only have experienced a reduced motivational spillover from being forced to observe a highly productive peer. The simulations also reveal that there is no evidence for a choice effect in our setting, i.e. workers do not get more motivated by peers that—all other things equal—they have chosen themselves.

In the last part of the paper, we benchmark the effects of social comparisons using three treatments that interact social comparisons with financial incentives for production (ENDO×$, EXOBEST×$, RANK×$). We show that our key findings are robust to the introduction of performance-based bonus payments, and that social comparisons and monetary incentives act as complements in our context. In addition, we also report the results of a follow-up experiment aimed at contrasting the effects of social and non-social comparisons. This experiment reveals that social comparisons generate much larger motivational effects than comparable non-social goals, thereby underscoring that the social dimension of peer comparisons is crucial for our results.

Overall, our results highlight that social comparisons can in principle be leveraged to boost productivity, but that policies aimed at raising output might also have unintended consequences such as, e.g., raising workers' stress. While these unintended consequences tend to be ignored, we argue that they should be monitored more systematically as they might ultimately also affect the firms' overall performance.

The remainder of the paper is structured as follows: In the next section, we discuss our contributions to the literature. In Section 3, we provide details on our experimental design and on our sample. We present our results in Section 4, and we conclude in Section 5.

2 Related Literature

Our paper relates to multiple strands of the literature and makes several contributions. First and most directly, our paper relates to the literature on the effects of social comparisons on productivity.[6] Existing studies have focused on the effects of relative performance feedback (see e.g. Charness et al., 2013; Gill et al., 2019) and randomly assigned peers on productivity (see e.g. Sacerdote, 2001; Falk and Ichino, 2006; Mas

[6]While we only focus here on the literature on social comparisons and productivity, note that an growing body of research investigates the effects of pay comparisons and salary transparency initiatives (see e.g. Perez-Truglia, 2020; Cullen and Perez-Truglia, 2022; Card et al., 2012).

and Moretti, 2009; Bandiera et al., 2010; Beugnot et al., 2019).[7] We contribute to this literature by providing new causal evidence that social comparisons can be leveraged to boost productivity, by highlighting the central role of peer assignment mechanisms and their interactions with monetary incentives, and by showing that social comparisons can have a substantial effect on workers' experienced stress. We also contribute to this literature by distinguishing the effects of social comparisons with those of non-social goals (see e.g. Corgnet et al., 2015, 2018).[8] In addition, our paper differs from previous research in that it isolates the behavioral effects arising from observing others, shutting down alternative channels such as, e.g., the effects of being observed, productivity complementarities, or social learning; thereby allowing us to isolate motivational spillovers. In this regard, we also make a methodological contribution by developing a novel experimental paradigm that permits the identification of social spillovers while circumventing the main hurdles pertaining their estimation (Manski, 1993).

Our paper also relates to the growing literature interested in designing non-random peer assignment mechanisms. Recent papers have theorized that exogenous peer-assignment rules that maximize productivity could be engineered (Graham et al., 2014; Roels and Su, 2014; Kräkel, 2016). However, empirical evidence on this conjecture is scarce, inconclusive, and limited to the context of education (Carrell et al., 2013; Chen and Gong, 2018). To our knowledge, our paper is the first to implement and assess the behavioral effects of such a policy in the context of a labor intensive assignment without complementarities, and to contrast it with alternative peer assignment mechanisms. Our findings show that forcing workers to compare to the predicted most motivating peer boosts productivity, but also generates the largest increase in stress. In contrast, letting workers choose whom to compare to is an effective and easily implementable way to boost productivity that does not have the disadvantage of generating such a large increase in stress. While outside of the scope of this paper, these results suggest that different peer assignment mechanisms may have different welfare implications—consistent with the recent discussion on the welfare effects of "social nudges" (Allcott and Kessler, 2019; Butera et al., 2022).

Perhaps most closely related to this strand of the literature and to our paper is Kiessling et al. (2021), who compare the effects of self-selected and randomly assigned peers using a framed field experiment in the context of a running contest organized

[7] For a review on the effects of performance feedback and peer effects at the workplace and in the laboratory, see Villeval (2020). For a review of social incentives in organizations, see Ashraf and Bandiera (2018).

[8] Throughout the paper, we use the terminology "non-social" to refer to any comparison that is made with a non-human "reference point", irrespective of whether or not the reference point has been set by a human being (e.g. the worker's superior).

at school.[9] In addition to an obvious change of context, our study differs from theirs in at least four important aspects. First, we assess the effects of social comparisons not only for performance but also for experienced stress while they focus only on the former. The inclusion of this second dimension allows us to better differentiate the implications of different assignment procedures, because we can illustrate that policies with similar performance effects may differ substantially in their unintended consequences. Second, we implement and assess the effects of a substantially larger number of assignment policies. Instead of focusing mainly on the comparison between random assignment and endogenous choice, we also implement an exogenous peer-assignment policy aimed at maximizing performance, and we interact all our central social treatments with financial incentives. These additional treatments not only enable us to cleanly disentangle the different behavioral mechanisms that distinguish endogenous from exogenous comparisons, but also make it possible to benchmark our effect sizes and to demonstrate their robustness to changes in workers' compensation scheme. Third, their setup involves simultaneous interactions between pupils, which raises the question of a possible reflection problem (Manski, 1993), and does not control for potential feedback effects since participants remain uninformed about their performance until the end of their experiment. Our experimental design, in contrast, builds on comparisons with a pre-determined reference population which allows us to cleanly account for both of these identification challenges. Fourth, we show that the social dimension of peer comparisons is a key driver of our findings: social comparisons generate much larger motivational effects than comparable non-social goals.

More generally, our paper is connected to the growing literature interested in understanding how individuals are influenced by their social environment. Social spillovers typically involve two simultaneous forces: the effects of observing and being observed by others. Recently, a growing literature has investigated the role played by the latter, for example in the context of education (Bursztyn and Jensen, 2015; Bursztyn et al., 2019) or voting (Gerber et al., 2008; DellaVigna et al., 2016).[10] This literature has typically relied on exogenously manipulating whether an individual's actions are obsevable to others or not. Our experimental paradigm can be considered the flipside of these studies, as we are exogenously varying *what individuals observe*, shutting down the "being observed" channel. While we focus on the effects of social comparisons in the context of effort provision, our experimental paradigm could

[9] See also (Falk and Knell, 2004) who present a simple theoretical framework for endogenous choice of social reference points.

[10] For a recent review on the literature on social pressure, social image and self-image, i.e. the effects of being observed, see Bursztyn and Jensen (2017).

easily be applied to other contexts where "observing others" is believed to be an important driver of behavior.[11] In this context, another important empirical question is what constitutes a relevant social reference group, i.e. to whom do people compare to, and why? Only a few recent studies have started to explore these questions (see e.g. Clark and Senik, 2010; Cicala et al., 2018; Kiessling et al., 2019), predominantly in the educational context. Our findings contribute to this discussion by providing new empirical evidence on preferences for peers in the context of a labour intensive task, and by exploring their underlying motives.

3 Experimental Design

Figure 1 provides an overview of the experimental design and of the different treatments. Our study comprises two sets of participants: participants who form our 'reference population' (left most column) and participants who took part in the main experiment (columns 2 to 8). All our participants are required to work on a real-effort task in two consecutive periods ('Effort 1' and 'Effort 2'), for which they are paid a fixed wage.

We first collected data on the reference population. For these participants, the experiment merely consisted of these two rounds of effort provision during which they "only" receive real-time feedback about their own production output. As we explain below, these participants constituted the relevant (social) environment for all the remaining participants.

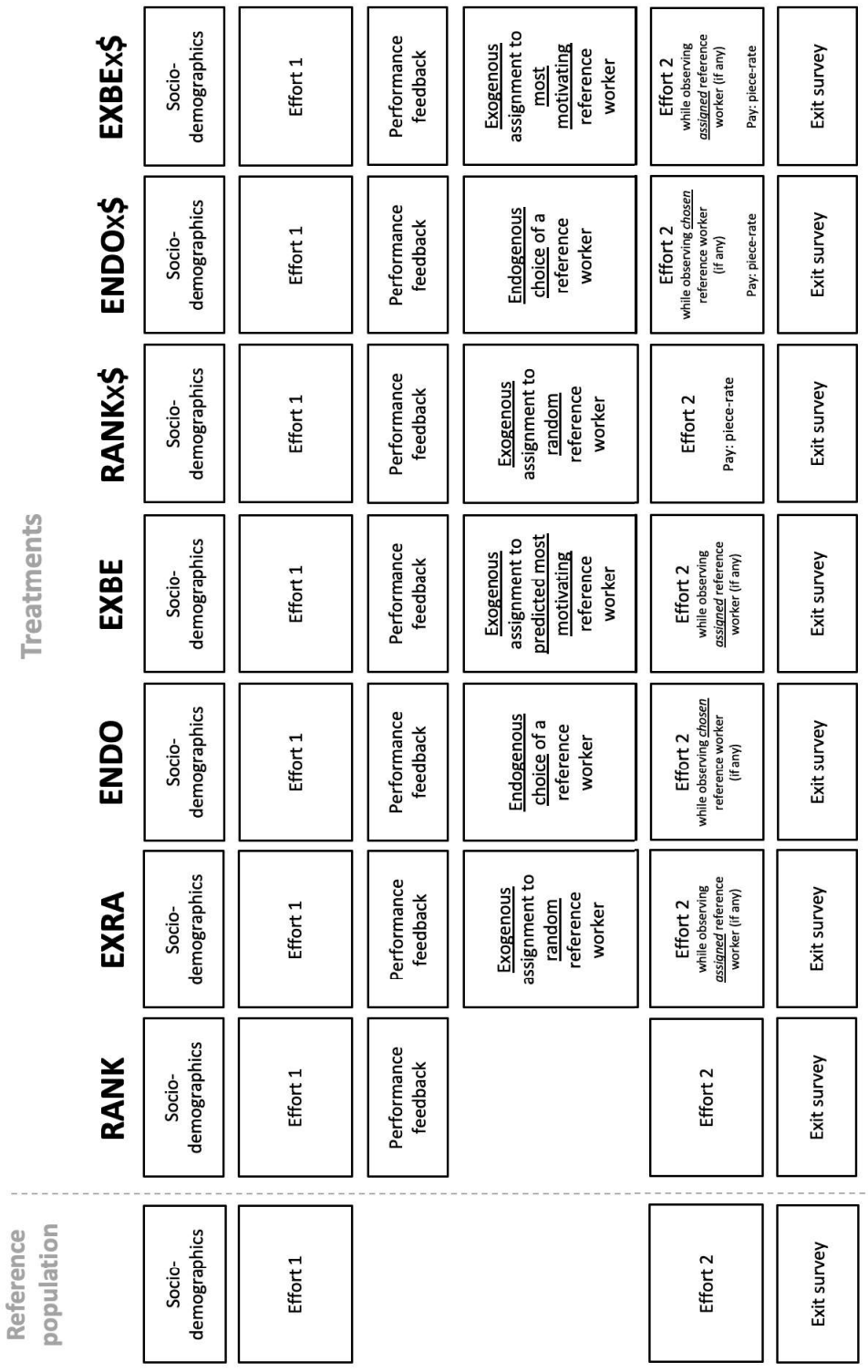
Shortly after collecting the data for the reference population, we collected data for the main experiment (which is the main focus of our study). For these participants, the experiment involved additional steps. They also started with a first round of effort provision (Effort 1) during which they also received real-time feedback about their own production output. Upon completion of this first round, they *privately* learn how their productivity in round 1 compares with the productivity in round 1 of the workers from the reference population ('Performance feedback'). They are then randomized into different treatments that vary whether and how they are exposed to real-time information about the round 2 performance of a reference worker (who is drawn from the reference population) while they are themselves working on the task

[11] In this spirit, recent papers have shown that providing individuals with social information about the behavior of others can affect behavior in variety of settings such as the provision of public goods (Chen et al., 2010), financial decision making (Bursztyn et al., 2014; Kirchler et al., 2018; Schwerter, 2019), labor market decisions (Coffman et al., 2017), and energy consumption (see e.g. Allcott and Kessler, 2019), among others. While these studies typically provide *static* or *aggregate* information (e.g. the average behavior of neighbors) to their subjects, we provide *individualized, real-time* information about the behavior of peers.

for a second time ('Effort 2').

In the following, we provide more details on the real-effort task, the different treatments, the reference population and the reference workers, as well as the sample.

Figure 1: Overview of the experimental design



Note: Our experimental design comprises two set of participants. The reference population (column 1), whose performance at the real effort task was measured in two consecutive rounds, serves as a source of (social) information to be provided to the main participants. The main participants are randomized (between-subjects) into one of seven treatments (columns 2-8). The treatments vary i) whether participants have the opportunity to observe the real-time work progress of a peer (a "reference worker") while they are completing the task in period 2, ii) how participants are matched with a reference worker (random assignment to a reference worker, targeted assignment to the predicted most motivating reference worker, or endogenous choice of a reference worker), and iii) the compensation scheme used to pay participants in period 2 (fixed wage vs. performance pay)

3.1 The Real Effort Task

As a basis for our experiment, we searched for a task with the following characteristics: i) the task requires real effort from workers, ii) the task generates substantial productivity differences across individuals, so that workers have a meaningful choice when choosing whom to compare to, iii) real-time comparisons between workers need to be simple and salient, so that they can have an impact on workers, and iv) observing another worker cannot allow an individual to get better at the task, so that motivational spillovers are not confounded with social learning.

The so-called a-b-task (Amir and Ariely, 2008; Berger and Pope, 2011; DellaVigna and Pope, 2017; Butera et al., 2022) fulfills all the above mentioned requirements. The task consists of alternatively pressing the 'a' and 'b' keys on a computer keyboard. Each a-b sequence adds a unit to the participant's output measure. Workers are instructed to produce as many units of output as possible while working on the task for 5 minutes in each period of the study.

Although the task is abstract and lacks intrinsic meaning, it shares the main characteristics of typical clerical and manual jobs (for a longer discussion, see DellaVigna and Pope, 2017). Most importantly, it is effort-intensive, repetitive, and tiring. These features also characterize the typical simple jobs that are often studied in field studies on worker motivation, such as fruit-picking (see e.g. Bandiera et al., 2005, 2010), tree-planting (Shearer, 2004), windshield installing (Lazear, 2000), or data-entry jobs (Kube et al., 2012, 2013).

3.2 Random Assignment of Reference Workers (EXRA)

We first describe the details of the treatment in which workers are exogenously (and randomly) assigned to reference workers (EXRA).[12] We provide a detailed description of the remaining treatments, and their key differences with EXRA, in Section 3.4.

Before participants started the experiment, they were informed that they would be paid a fixed wage for their participation.[13] They were made aware that the study would consist of several parts, and that they would receive instructions separately

[12]The relevant screenshots are all provided in Appendix A.1.1.

[13]We purposefully chose a fixed wage in order to be able to cleanly disentangle motivational spillovers from alternative mechanisms that might be at play when individuals compare themselves with others under a pay-for-performance contract. For example, in the presence of a piece rate the effects of motivational spillovers would be confounded with the effects of pay differentials. Fixed-wage contracts are empirically relevant as a substantial share of the workforce is compensated with such contracts, and they do not prevent workers from exerting effort (see e.g. DellaVigna and Pope, 2017). For a longer discussion of the benefits of using fixed-wages in a related context, see Charness et al. (2013). For a comparison of the effects of fixed wages, piece-rates and non-monetary incentives in the a-b-task, see DellaVigna and Pope (2017).

for each part. Participants were therefore only informed about the part of the experiment that they were about to complete and were unaware of what would come next. This feature of the design implies that participants were not aware of any treatment-specific details when completing the first work period. Period-1 performance is therefore fully comparable across treatments and can be used as a clean measure of a participant's baseline productivity.

Part 1: Socio-Demographics The experiment started with questions on participants' socio-demographics (see Appendix F.1 for details).

Part 2: Production Period 1 (a-b Task) Upon completion of the questionnaire, participants received instructions for the a-b task. The instructions emphasized that their task was to sequentially press the "a" an "b" buttons as quickly as possible during 5 minutes. Participants went through a practice round of 15 seconds to familiarize themselves with the task. They were then asked to give an estimate of how many points they thought they would be able to reach, and were then asked to work on the task for 5 minutes. While working on the task, participants were constantly updated on their current output (both numerically and graphically using a growing vertical bar) and the remaining time (see the screenshot provided in Figure A.1 in the Appendix). Upon completion of the task, we elicited participants' stress levels. We also asked whether they were satisfied with their performance, and whether they found the task difficult (see Appendix F.2 for details).

Part 3: Performance feedback Upon completion of period 1, participants learned that they would have to complete the a-b task a second time. However, they were informed that their performance would first be compared to the performance of 60 other participants who had completed the exact same task at an earlier point in time (we provide more about this "reference population" below). The instructions emphasized that the only aim of this ranking was to provide them with information about their performance, that it was private information (i.e. that it would never be visible to anyone else but the participant), and that it had no influence on their payment. Participants were then shown a table displaying their own performance, along with the performance of all 60 workers in the reference population. To facilitate comparisons, participant's own position within this table was highlighted (see screenshot A.2 in Appendix for details).

Part 4: Random Assignment to a Reference Worker Participants were then informed that the computer might assign them to one of three workers from the reference population and they were reminded of the first round performance of these workers (we provide more details regarding these "reference workers" in the next subsection). Participants learned that—if matched with a worker—they would get to observe the evolution of this other worker's performance in round 2 while working on the task. They were also made aware that computer might not assign them to another participant, in which case they would complete round 2 in the same conditions as round 1. A screenshot of this stage is provided in Figure A.3 in the Appendix.

Part 5: Production Period 2 (a-b Task) The second work period was organized in the same way as the first one, with the exception that participants who were assigned to a reference worker could now constantly compare their output to the evolution of the second round output of that worker in real time. This new piece of information was displayed to them both numerically and as a growing vertical bar (see figure A.4 in Appendix). Participants who were *not* assigned to a reference worker completed the second round in the exact same conditions as in the first round. At the end of this second round, participants were informed about their output and were again invited to report their stress level.

Part 6: Exit survey and profit information Before exiting the study, participants were asked to fill out a short questionnaire aimed at eliciting their perceptions of the reference worker and its effects (see Appendix F.3 for details). They were then informed about their profit and the payment procedure.

3.3 Reference Population and Reference Workers

Shortly before launching our main study, we collected data on 60 workers (see 'Reference Population' in Figure 1). These workers completed a version of the experiment in which they only completed the real effort task twice, but only received feedback about their own performance. The workers of the reference population never received any information about other workers (i.e., Parts 3 and 4 of the EXRA treatment described in Section 3.2 were skipped). This reference population constitutes the relevant social environment for all participants in our experiment in the sense that *all* the social information with which participants were confronted came from this group. Indeed, when participants were informed about their performance in round 1 of the a-b task, they learned how their output compared to the output in round 1 of these

60 workers. Moreover, the potential reference workers to whom participants might have been assigned in round 2 were also selected from this reference population.

To ensure a reasonable level of statistical power and to have a sufficient number of observations per reference workers, we restricted the set of potential reference workers to three individuals. We selected these three workers based on the following criteria: a) the workers needed to differ substantially in terms of their performance at the task in both rounds, b) they had to improve by about 10% between period 1 and period 2, and c) they were required to have a relatively constant production output throughout each round.[14] These criteria led us to select the following three workers: a high productivity worker (HI, worker ranked 4 out of the 60 workers from the reference population), an average productivity worker (MI, ranked 26 out of 60) and a low productivity worker (LO, ranked 49 out of 60).[15]

Using a fixed and pre-determined set of workers (who completed the study prior to the main experiment) as a reference population is a central feature of our design. It allows us to circumvent two problems that often plague (observational) studies on social comparisons. First, it allows to circumvent the reflection problem (Manski, 1993) that arises when trying to identify social spillovers in the context of simultaneous interactions, i.e. when the observing and the observed participant affect each other at the same time. We circumvent this problem by providing workers with information about the output of reference workers that are drawn from the reference population and that can therefore no longer change their behavior as a response to being observed. Second, our design circumvents the feedback problem, which refers to the fact that observing a peer may not only impact performance through motivational spillovers but also through an informational channel (see e.g. Charness et al., 2013). We control for this important confound by providing workers from all the treatments with performance feedback after period 1—thereby allowing us to cleanly isolate motivational spillovers and distinguish them from informational effects. Here too, the use of a fixed reference population has an advantage: It ensures that the performance feedback is constant across both treatments and participants.

[14]Reference workers with erratic patterns (e.g. working extremely fast for the first 2 minutes, doing nothing for one minute, and then working fast again during the last two minutes) might have confounded comparisons, because they would not have differed only in the performance dimension, but also in the way in which they completed the task.

[15]We provide the full distribution of the performance of our reference population in both periods in Appendix A.2. In addition, we also depict the production paths of the three potential reference workers in both rounds (see Figures A.6 to A.8).

3.4 Additional Treatments

We use EXRA as our baseline condition to establish the causal effects of social comparisons. Because this treatment randomly assigns workers to reference workers, it allows us to identify the causal effects of observing a peer and to assess whether different reference workers (LO, MI and HI) affect the workforce differently. To account for possible feedback effects, we compare the EXRA condition to the RANK condition described below.

Rank Information Only (RANK) In RANK, participants *only* received information about how they ranked compared to the 60 workers from the reference population. However, they were *not* told anything about the reference workers and they did *not* get any feedback about the performance of another participant as they completed round 2 of the a-b task.

We explore the extent to which non-random peer assignment procedures can be used to leverage the productivity impact of social incentives by comparing the effects of randomly assigned reference workers (EXRA) to those of endogenously chosen reference workers (ENDO) and those of a policy that exogenously assigns workers to their predicted most motivating reference worker (EXBE).

Endogenous Choice of Reference Workers (ENDO) This treatment is very similar to EXRA, with the exception that participants were given the opportunity to *choose* their reference worker. To keep things as comparable as possible, the way in which reference workers were introduced remained identical to the EXRA condition and the choice set included the exact same three reference workers (see the screenshot in A.5 in Appendix A.1.2). Participants could also decide *not* to observe any reference worker. ENDO is fundamentally different from our other treatments in that the analysis of its effects not only requires investigating the motivational spillovers from the reference worker on the participant, but also necessitates an examination of the choice process. To gain further insights into participants choice motives, we also asked them to motivate their choice in an open-text format.

Exogenous Assignment to Predicted "Best" Reference Workers (EXBE) In this treatment, participants were exogenously assigned to the reference worker that was predicted to have the largest positive effect on their performance on the basis of the data collected in EXRA, i.e., participants in EXBE were matched with their *predicted* most motivating reference worker based on their observable characteristics. Impor-

tantly, the wording was kept exactly identical to the one used in EXRA, i.e. participants did *not* know that they would be assigned to the reference worker that is predicted to maximize their productivity (see the screenshot of the EXRA treatment provided in Figure A.3 in the Appendix). We provide more details on the tailoring of this matching procedure in Section 4.2.

In addition, we also implemented three additional treatments in which participants were compensated for performance (on top of their fixed payment). We use these treatments to benchmark the effects of social comparisons and to test their robustness to changes in the compensation scheme.

Incentives (RANK$, ENDO$, EXBE$) These three treatments were identical to RANK, ENDO, and EXBE, but participants were paid a piece-rate in addition to the fixed wage that workers received in all other treatments. We provide more detailed information on the level of the piece rate and its impact on compensation when we report the results of these treatments in section 4.4.

We summarize the main features of each treatment and report the respective sample sizes in Table A.2 in Appendix. We collected data on all these treatments simultaneously, with the exception of EXRA that had to be run slightly ahead of time in order to be able to tailor the EXBE condition. To control for possible unexpected changes in the subject pool within these few days, we collected half of the RANK data together with EXRA and the other half with the rest of the treatments. As we discuss below, we find no significant differences across these two waves of data collection.

3.5 Sample and experimental protocol

We pre-registered our study on the AEA RCT Registry (AEARCTR-0003217).[16] We ran our study on Amazon Mechanical Turk (MTurk), an online labor market where employers can advertise small jobs ('HITs') that typically consist of simple, repetitive tasks.[17] Workers ('MTurkers') can complete any HIT they like, provided that they fulfill the enrollment criteria defined by the employer. Because the platform allows to assign a large set of small tasks to a very large set of workers in a short

[16]While our experiment was conducted exactly as pre-registered and our analysis largely follows the pre-analysis plan, we slightly deviate from the pre-analysis plan on occasions. For transparency, we discuss these deviations and provide the interested reader with a populated pre-analysis plan (Banerjee et al., 2020) in the online Appendix.

[17]Examples of typical tasks assigned to MTurkers include encoding text depicted on a picture, rating the quality of short audio recordings, or assessing the emotional-state of photographed individuals.

amount of time, it is no surprise that it is being increasingly used by academic researchers, including economists, to conduct large-scale between-subjects studies (see e.g. DellaVigna and Pope, 2017; De Quidt et al., 2018; Almås et al., 2020; Cappelen et al., 2021, 2023).[18] For example, DellaVigna and Pope (2017) also use MTurk and the a-b task to investigate the effects of different financial and non financial incentive schemes on effort.

We conducted our experiment over a period of about 2 weeks in mid-August 2018. The experiment took about 10 minutes to complete, for which we paid a fixed wage of \$1.5.[19] We required that workers are US residents, that they have an approval rate of at least 95%, and a minimum of 50 approved tasks. In addition, our experimental protocol prevented individuals from taking the same HIT twice. Eligible MTurkers were automatically redirected to our own server, and randomized into a treatment (between-subjects). An important feature of our design is that the different treatments are implemented in the second half of the study, i.e. absolutely everything that the workers see during the first half of the study (including the HIT description) is the same across all the treatments. This prevents that workers with different characteristics select into different treatments, and substantially limits the odds that attrition differs by condition.

In total, 6635 eligible workers completed our HIT. From these, we excluded (i) workers who scored more than 2000 points per round[20], (ii) workers who exited and re-entered the task, and (iii) workers who did not complete the entire study within 60 minutes of starting. These sample restrictions were all pre-registered. In addition, we also excluded a few workers who incurred technical problems with our study.[21]

[18]While questions about the generalizability of experimental findings may arise, evidence from a recent meta-study indicates that peer effects estimated in the context of laboratory studies generalize to the field (Herbst and Mas, 2015). In addition, recent comparative studies find no substantial differences between findings documented using MTurk and findings documented in alternative samples (see eg. Horton et al., 2011; Snowberg and Yariv, 2021). Moreover, while worries about subject pool representativity, innatention, and bots can be legitimate in some settings (e.g. when studying political preferences), they are unlikely to matter in our study since workers' only task is to *exert* real effort at a task that would be very difficult to automate.

[19]We recruited workers by advertising a "study on decision making." We set up the incentives such that the assignment would be attractive for workers. In particular, we made sure that participants would earn more than the US legal minimum wage of USD 7.5/hour independent of their productivity. Note that in some treatments (RANK×\$, ENDO×\$, EXBE×\$) workers also received a piece rate in addition to this flat wage, as we will discuss in Section 4.4.

[20]Results from our own pilots and the pilots run by DellaVigna and Pope (2017) suggest that it is virtually impossible to score more than 2000 points within 5 minutes without cheating at the task.

[21]Most of these workers sent us emails mentioning that the program would not keep track of their score at the 'a-b' task, i.e. despite clicking on 'a-b', i.e. their total output always remained equal to zero. This problem was also faced by some subjects in DellaVigna and Pope (2017). Importantly, this additional restriction is immaterial to our results.

The final sample includes 6532 subjects.[22]

We display the main summary statistics of our sample in Table A.3 in the Appendix. 40 percent of our sample are male participants, and the average worker is 36.2 years old. The average exerted effort in round 1 is 1042.2, and it is 1173.1 in round 2. In Table A.4 in the Appendix, we show that workers characteristics are well balanced across the different treatments. In particular, productivity in round 1 (effort 1) is orthogonal to the treatments. Finally, we also show that attrition is unrelated to the treatments (see Table A.5 in Appendix).

4 Results

We present our results in several steps. First, we establish the causal effects of randomly assigned reference workers (EXRA) and contrast them with those of rank information (RANK). Second, we assess the effects of letting workers choose whom to compare to (ENDO), and those of a targeted exogenous assignment policy aimed at maximizing motivational spillovers (EXBE). We then explore the behavioral mechanisms that distinguish endogenous and exogenous comparisons. Third, we benchmark the effect sizes of social comparisons and establish their robustness using treatments that interact social information with monetary incentives for production (RANK×$, ENDO×$, EXBE×$). Finally, we also contrast the motivational effects of social comparisons with those of non-social goals.

4.1 The causal effects of randomly assigned reference workers

4.1.1 Average treatment effects

We start our analysis by establishing the motivational effects of randomly assigned reference workers. When comparing themselves to a reference worker, participants naturally gather information about their (relative) productivity. This information might in itself affect participants. Indeed, previous work has shown that providing workers with ranking information alone can induce substantial increases productivity, even in situations where the incentives to achieve a better rank are entirely symbolic (see e.g. Charness et al., 2013). In order to properly disentangle motivational

[22]We pre-registered samples of 500 subjects in treatments where participants *cannot* choose their reference worker, and 1000 subjects in treatments where subjects are given the possibility to choose their reference worker. We doubled the sample size in the treatments with endogenous choice because we expected a lot of between-subject heterogeneity. This allows us to reach higher precision when analyzing the behavior of workers, conditional on the reference worker they chose. We display the exact number of observations per treatment in Table A.2 in the Appendix A.3.

spillovers from such feedback effects, we compare participants in EXRA with those in RANK.

It is not obvious that randomly assigned reference workers have an overall positive performance effect at the treatment level. On one hand, comparisons with a reference worker may strengthen participants' desire to compete, or motivate them to "keep on working hard" when they would rather slow down or give up. On the other hand, however, observing a randomly assigned reference worker might also be demotivating, especially if the assigned reference worker is unproductive (so that being more productive than the reference worker requires only little effort) or very productive (so that catching up with the reference worker seems impossible and becomes frustrating).

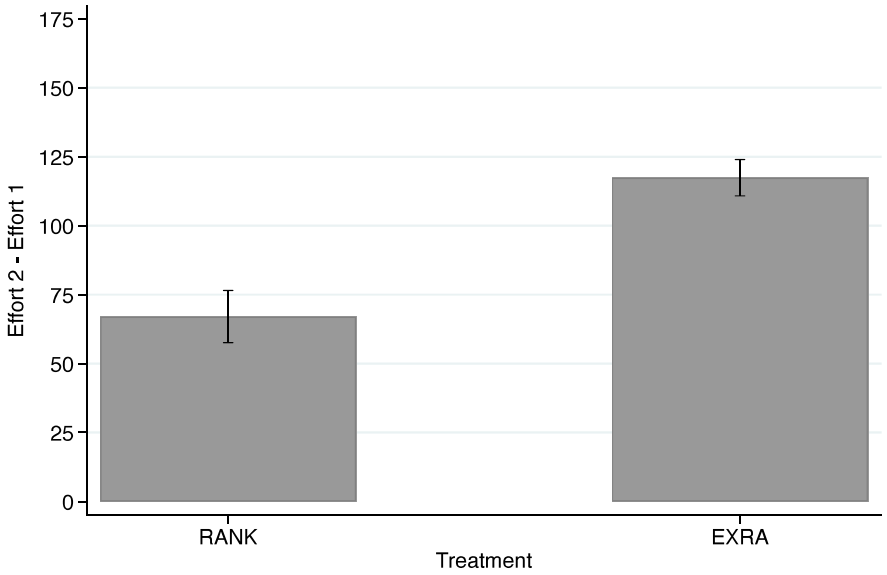
Figure 2 depicts the average increase in performance between period 1 and period 2 (effort2-effort1) in RANK and EXRA.[23] The provision of information about relative performance in RANK increases production by 67 units (from 1040.6 in period 1 to 1107.6 in period 2, an increase of 6.5 percent, $p < 0.01$).[24,25] Exposing workers to randomly assigned reference workers generates an average increase in performance that goes far beyond this pure information effect. Indeed, the performance of workers in EXRA increases by 117 units (from 1039.9 in period 1 to 1157.3 in period 2, an increase of 11.3 percent, $p < 0.01$), i.e. about *twice* the size of the increase in RANK (this diff-in-diff comparision is significant, $p < 0.01$). Thus, assigning participants to randomly drawn reference workers generates—on average—motivational spillovers that exceed those obtained by the sole provision of rank information.

[23] In order to be able to cleanly compare RANK with the other treatments and because EXRA was collected slightly before the remaining treatments, we collected a sample of 500 participants in RANK in *both* waves (see Section 3.5 for details). There is no statistically significant difference in period 1 performance between these two samples (Wave 1: 1027 units, Wave 2: 1053 units, $p = 0.16$) and we therefore pool them together. This procedure is immaterial to our results.

[24] All the p-values reported in this paper are based on Wald tests of linear hypotheses about the parameters of OLS estimations in which we regress the dependent variable on treatment dummies and interactions between treatment dummies and an indicator variable for period 2 (since treatments are operationalized in period 2). An advantage of this procedure is that it also allows to control for a workers' individual characteristics. Without controls, the p-vales obtained are equivalent to those obtained using two-samples t-tests. For more details on the estimation procedure and for the regression outputs, see Appendix B2.

[25] In principle, the change in effort between rounds can be explained by a combination of learning effects and treatment-specific features. Because we are comparing changes in effort *across* treatments, our design is well suited to isolate the effects of relative performance feedback and social comparisons, holding learning effects constant. Moreover, data from the reference population suggests that the change in effort that can be attributed to learning is small and insignificant ($p = 0.40$).

Figure 2: Average increase in performance for workers who only see their rank (RANK) and for workers who are in the random assignment condition (EXRA)



Note: The figure depicts the average increase in performance from round 1 to round 2 in RANK and EXRA. In RANK, workers are only informed about their ranking before proceeding to round 2. In EXRA, workers are informed about their rank and are then randomly assigned to either one of the reference workers (LO, MI, HI) or to no reference worker. Whiskers represent the standard errors.

4.1.2 What are the effects of different reference workers?

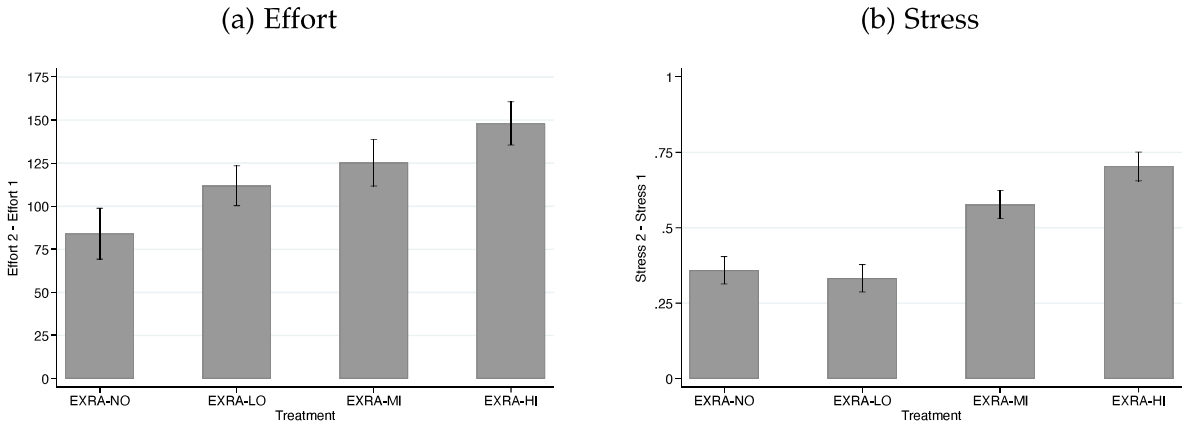
To better understand the performance enhancing effect of randomly assigned reference workers, we explore how our three different reference workers affect participants' productivity. Figure 3a displays the average performance increase for each of the four sub-conditions of EXRA: no reference worker (EXRA-NO), low-productivity reference worker (EXRA-LO), medium-productivity reference worker (EXRA-MI), and high-productivity reference worker (EXRA-HI).[26]

A very clear pattern emerges from this figure: On average, the performance increase from period 1 to period 2 becomes larger as the productivity of the assigned reference worker increases. Participants who were exogenously assigned to the least productive reference worker improve by 111 units (from 1051.2 in period 1 to 1163.1 in period 2, an increase of 10.6 percent, $p < 0.01$), while those who were assigned to the medium-performance reference worker improve by 125 (from 1021.2 in period 1 to 1146.5 in period 2, an increase of 12.3 percent, $p < 0.01$) and those who were assigned to the most productive reference worker improve by 148 units (from 1044.8 in

[26]Workers in EXRA are randomly (and uniformly) assigned to one of the four treatment arms: or EXRA-NO, EXRA-LO, EXRA-MI, or EXRA-HI. We therefore have approximately 500 observations per treatment arm (see Appendix A.3 for details).

period 1 to 1192.8 in period 2, an increase of 14.2 percent, $p < 0.01$). All these increases in performance are significantly larger than the performance increase documented in RANK (each of the three diff-in-diff tests is significant, with $p < 0.01$).

Figure 3: The effects of the different randomly assigned reference workers



Note: Workers in EXRA were randomly (and uniformly) assigned to either one of the three potential reference workers (LO, MI, HI) or to no reference worker (NO). Each bar corresponds to one of the four treatment arms of EXRA. Panel a) depicts the average change in effort between rounds 1 and 2. Panel b) depicts the average change in stress between rounds 1 and 2. Stress levels were measured after each round using the question "On a scale from 1 to 5, how stressed have you been while completing the task?" Answer categories ranged from "Not at all stressed" (1) to "Very stressed" (5). Whiskers represent the standard errors.

Moreover, getting assigned to a more productive reference worker does, on average, generate a higher increase in performance than getting assigned to a less productive reference worker: Workers in EXRA-MI increase their performance slightly more than those in EXRA-LO, and workers in EXRA-HI increase their performance much more than workers in EXRA-LO ($p = 0.03$) and slightly more than those in EXRA-MI (although this test is insignificant).

Finally, consistent with our findings in RANK, participants who were not assigned to any reference worker improve their performance by only 84 units (from 1043 in period 1 to 1127.1 in period 2, an increase of 8 percent, $p < 0.01$), which is not significantly different from the increase in performance documented in RANK ($p = 0.33$).

So far, we have demonstrated that randomly assigned reference workers generate substantial motivational spillovers, but recent evidence from psychology suggests that social comparisons might also affect well-being (see e.g. Buunk and Dijkstra, 2017; Bárcena-Martín et al., 2017; Fujita and Diener, 1997). In our context, social comparisons might affect the levels of stress experienced by our participants. We therefore depict the average increase in stress reported by our participants in the

different treatment arms of EXRA in Figure 3b.[27,28] The figure shows that subjects assigned to no reference worker or to the least productive reference worker experience a significant but only very moderate increases in stress (+0.36 in EXRA-NO and +0.33 in EXRA-LO, which correspond to increases of 28 and 26 percent of a standard deviation, respectively, both $p < 0.01$). In contrast, participants assigned to EXRA-MI report an increase in stress of +0.58 (i.e. +45.6 percent of a standard deviation) which is a significantly larger difference compared to EXRA-LO ($p < 0.01$) and workers in EXRA-HI report an even larger increase in stress of +0.70 (i.e. +55 percent of a standard deviation, $p < 0.01$), which is significantly different than the increase reported in the remaining treatments.[29] Thus, observing a more productive reference worker not only increases productivity, but it also substantially raises the levels of stress experienced by the participants.[30]

4.1.3 How do these effects depend on the characteristics of the observer?

Whereas the above analysis suggests that, overall, more productive reference workers are more motivating, there are plausible reasons to expect that this result might not hold for all participants. For example, some workers might get discouraged if they are assigned to a peer who is so much more productive that it seems impossible for them to catch up. For such participants, being assigned to a reference worker whose performance is only slightly better than their own might be more motivating. Other workers' performance might be highest if they work in isolation and do *not* observe

[27]Following recent papers in economics and psychology (see e.g. Haushofer and Shapiro, 2016; Haushofer et al., 2015, 2021; Esopo et al., 2019), we measured perceived stress using a self-reported question. An advantage of using a single-item question is that it permits to measure stress in a more obfuscated way (by "hiding it" around two unrelated questions—like we did) than using a battery of questions. Specifically, stress levels were measured after each round using the question "On a scale from 1 to 5, how stressed have you been while completing the task?" where 1 means "Not at all stressed" and 5 means "Very stressed." The average level of stress after period 1 is 2.37, with a standard deviation of 1.27. The increases in stress reported throughout the paper are expressed in relation to this standard deviation.

[28]While we had anticipated that social comparisons might affect stress—which is why we included it in our study and in our pre-registration—we had no specific ex-ante hypotheses on these effects. In addition, we also collected data on two other variables (satisfaction about own performance and perceived task difficulty). For completeness, we also report the effects of our treatments on these other two variables in the Appendix E. As we explain therein, the findings are entirely consistent with the effects that social comparisons have on performance and on perceived stress.

[29]The increase in stress reported by participants in EXRA-HI is significantly larger than the one reported by participants in EXRA-MI ($p = 0.06$), EXRA-MI ($p < 0.01$) and EXRA-LO ($p < 0.01$).

[30] In the exit questionnaire, we also asked subjects to indicate the extent to which observing the reference worker made them nervous. The correlation between stress and nervousness is positive and highly significant ($\rho = 0.57$, $p < 0.01$), consistent with our interpretation of stress being a rather negative experience. While we do not display these results here due to space constraints, the treatment effects on nervousness are largely consistent with those on stress as well as those discussed in Appendix E.

any reference worker at all. This might be the case if, for example, a worker feels distracted by having to look at a reference worker, or if such comparisons are too stressful.

Interestingly, our data reveal that these effects do not seem to be of substantial importance in our context. In fact, we find that for virtually all workers, being exogenously assigned to the most productive reference worker (HI) generates the largest increase in productivity. The only exception are those workers whose performance in period 1 was lower than the performance of the least productive reference worker. For these workers in the lowest segment of the performance distribution, being matched with the least productive reference worker is the most motivating.

In addition, we also show that—consistent with the aggregate findings documented above—workers who are randomly assigned to the HI reference worker are the ones who generally experience the highest increase in stress, irrespective of how productive they were in round 1.

All these results are largely independent of the observer's gender. We document these results in detail in Appendix B.2.3.

4.1.4 Central takeaways from EXRA

Overall, these first results highlight the potentially large effects that social comparisons can have. Even in situations in which reference workers are *randomly* assigned, they can generate large increases in productivity. However, for most individuals, being assigned to a motivating reference worker also means experiencing a substantially larger amount of stress.

The finding that the effects of social comparisons depend on the productivity of the assigned reference worker points to the relevance of the matching procedure and the conjecture that random assignment of peers does most likely not fully exploit the motivational potential of social reference points. In the next section, we investigate the effects of two alternative assignment mechanisms that might further enhance social spillovers: endogenous choice, where workers are given the opportunity to choose whom to compare to, and targeted exogenous assignment in which workers are exogenously assigned to their predicted most motivating peer. We then investigate the nature of these endogenous comparisons, and compare their effects to those of increased monetary incentives.

4.2 Leveraging social comparisons using non-random peer assignment policies

In this section we explore the behavioral impact of two non-random matching procedures. In ENDO, participants are given the possibility to decide which reference worker to compare to (if any) in the second period. In EXBE, workers are exogenously assigned to their *predicted* most motivating reference worker based on their productivity in round 1 and their gender.[31] Beyond their practical relevance, the comparison of these matching procedures is also interesting from a behavioral perspective because they involve potentially important trade-offs.

Letting workers choose has the advantage that nobody is "forced" to observe a peer against their will. As a consequence, the frustration of being (mis)matched to an undesirable reference worker, as well as stressful comparisons, can be avoided. In addition, workers might find it particularly motivating to observe a reference worker that they have picked themselves (the choice effect). The potential downside is, however, that workers might select their peers for reasons other than their motivational potential so that the performance-enhancing effect of endogenous choice might be limited.

Exogenously assigning workers to their predicted most motivating peer has the obvious benefit that the impact on performance can be expected to be strong. At the same time, the full motivational potential may not be reached if some workers feel frustrated about being forced to observe a peer they did not want to observe (mismatch effect). Finally, targeted matching also risks to increase the perceived stress level substantially because for most workers the high-productivity reference worker is predicted to be most motivating but also the most stressful peer to observe (as we discussed above).

We depict the average change in productivity for ENDO and EXBE (along with RANK and EXRA) in Figure 4a below. The figure unambiguously shows that both EXBE and ENDO generate productivity increases that are larger than the one documented in EXRA: while participants in EXRA improve by an average of 117 units (from 1039.9 in period 1 to 1157.3 in period 2, i.e. an increase of 11.3 percent, $p < 0.01$), participants in EXBE improve by 146 units on average (from 1026.5 in period 1 to 1172.48 in period 2, i.e. an increase of 14.2 percent, $p < 0.01$) and participants in

[31] To predict which reference worker is the most motivating for a particular worker, we use data from the 2000 workers in EXRA to obtain a point estimate of performance for each reference worker (and no reference worker). We determine these point estimates for workers who reached different levels of output in round 1 and for different genders. As discussed above, for most participants (except the least productive ones), the most productive reference worker (HI) is predicted to be most motivating. For details, see Appendix B.5.

ENDO improve by an average of 138 units (from 1059.3 in period 1 to 1197.12 in period 2, i.e. an increase of 13 percent, $p < 0.01$). The performance increases in EXBE and ENDO are not significantly different from each other ($p = 0.57$), but the effects of both these treatments are larger than the effect of EXRA (both diff-in-diff tests yield $p < 0.05$).

While ENDO and EXBE have similar effects on workers' productivity, Figure 4b shows that these two treatments affect workers' stress levels differently. Letting workers choose whom to compare to leads to a moderate increase in stress of 0.65 points (+ 53 percent of a standard deviation, $p < 0.01$), which is significantly larger than the increase in stress experienced by workers in EXRA (+ 0.49 points, i.e. + 38 percent of a standard deviation, $p < 0.01$) who are randomly assigned to reference workers ($p < 0.01$). In contrast, forcing them to compare to their predicted most motivating reference worker yields an even larger increase in stress (+0.78 points, approximately +0.61 percent of a standard deviations, $p < 0.01$), which is significantly larger than in the other treatments (EXBE vs ENDO: $p = 0.02$; EXBE vs. EXRA : $p < 0.01$; EXBE vs. RANK : $p < 0.01$).

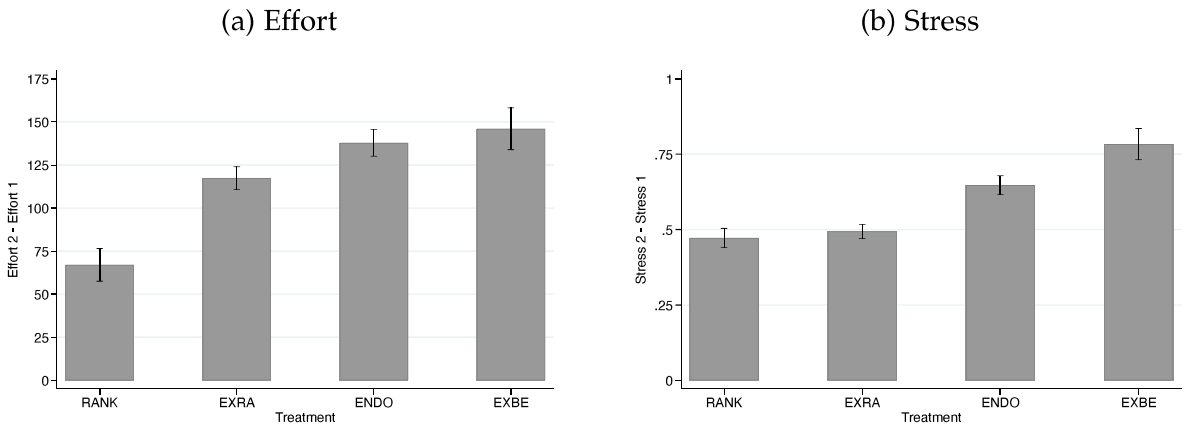
These results highlight the power of endogenous comparisons: letting workers choose whom to compare to generates a strong increases in productivity *without* increasing stress as much as assigning them to the predicted most motivating reference worker. This insight is interesting from a managerial perspective: in many real-life settings, implementing EXBE might be challenging and costly (because targeted matching requires detailed information about workers' predicted behavioral reactions to alternative peers). Our results suggest that—at least in certain settings—simply letting workers choose whom to compare to might be an attractive, and easier to implement alternative.

In what follows, we uncover the behavioral origins of these desirable effects of endogenous peer choice. To do so we analyze workers' preference for peers, we unveil the key motives that drive these choices, and we tease apart the different behavioral mechanisms that distinguish endogenous and exogenous social comparisons.

4.3 Endogenous social comparisons and motivational spillovers

Several behavioral channels that distinguish endogenous from exogenous comparisons may explain why ENDO produces a performance enhancing effect that is comparable to the one of EXBE without creating the same increase in perceived stress levels. First, there might be important differences between the set of reference workers that participants choose in ENDO and the set of references workers that workers

Figure 4: The effects of endogenously chosen reference workers and of targeted exogenous matching



Note: Panel a) depicts the average change in effort between rounds 1 and 2. Panel b) depicts the average change in stress between rounds 1 and 2. Stress levels were measured after each round using the question "On a scale from 1 to 5, how stressed have you been while completing the task?" Answer categories ranged from "Not at all stressed" (1) to "Very stressed" (5). Whiskers represent the standard errors.

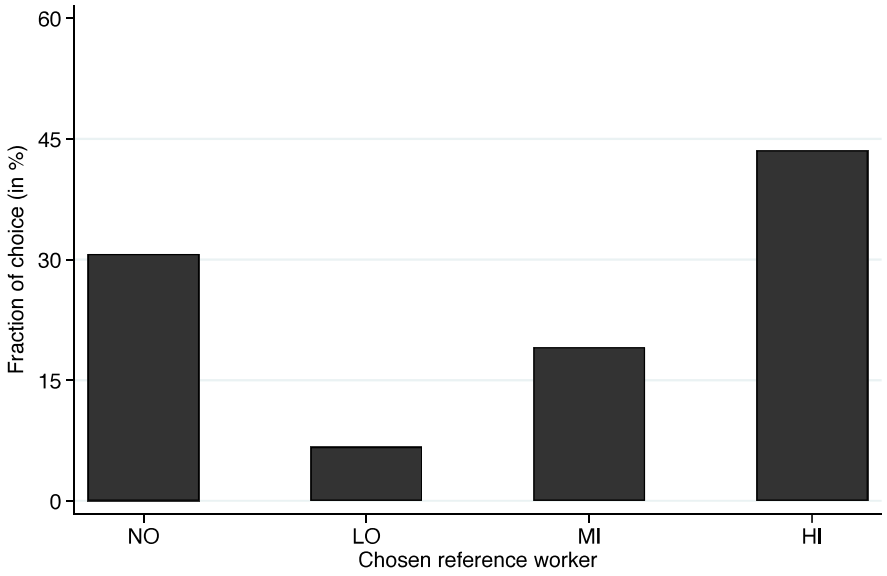
are assigned to in EXBE. Whereas workers in EXBE are predominantly assigned to the most productive reference worker, participants in ENDO might choose to compare to different reference workers. It is therefore important to understand participants' preferences for peers and their determinants. Second, exogenously assigning participants to reference workers implies that a substantial proportion of participants is matched with a reference worker they would not have chosen. Such mismatches might significantly reduce the motivational spillovers of social comparisons (the "*mismatch effect*)." Third, there might be a "*choice effect*", i.e. workers might be more strongly influenced, *ceteris paribus*, by a reference worker that they have chosen themselves. Such an effect might arise if participants pay *less* attention to exogenously assigned reference workers, or if they consider them *irrelevant*; so that the potential motivational spillovers of exogenously assigned reference workers are limited.

The richness of our experimental design enables us to isolate these different channels and to determine their relative importance. In the following, we first document participants' preferences for peers. In doing so, we not only explore the overall frequencies with which the different reference workers are chosen in ENDO, but we also investigate how participants' own characteristics such as their first-round productivity and their gender predict their choices. In addition, we explore workers' choice motives using text analysis. We then assess the relative importance of the mismatch effect and the choice effect using a series of simulations.

4.3.1 Preferences for peers and their determinants

Figure 5 depicts the relative frequency with which participants in ENDO choose each of the four available reference workers: no reference worker (NO), the weakly performing reference worker (LO), the intermediately performing reference worker (MI), and the strongly performing reference worker (HI). The most frequently chosen option is the best performing reference worker (43 percent), followed by the choice not to compare to any reference worker (31 percent). Interestingly, the other two alternatives (intermediately and weakly performing reference workers) are chosen less frequently (19 percent and 7 percent, respectively). A Pearson χ^2 test unambiguously rejects the null hypothesis that the different options are chosen equally often ($p < 0.001$), ruling out that participants either choose randomly or have uniformly distributed preferences.[32]

Figure 5: Distribution of chosen reference workers



Note: Distribution of choices of a reference worker in ENDO. 'NO' indicates the proportion of workers who choose *not* to compare to a reference worker. LO (MI, HI) indicates the proportion of workers who choose to compare to the weakest (average, strongest) reference worker.

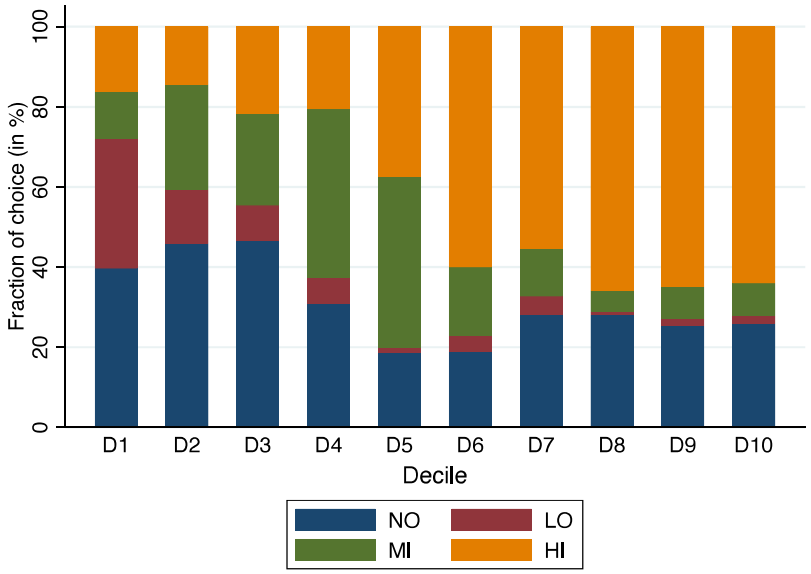
To shed light on the determinants of participants' choices, we now explore how their own productivity in period 1 (and their gender) affects their choices. In Figure 6, we depict the distribution of chosen reference workers as a function of the per-

[32]One might wonder whether participants' are able to predict the performance of the different reference workers, and their effects on their own performance. Our data on participants' beliefs suggests that they correctly anticipate the relative performance of the different reference workers in period 2, but they slightly underestimate the absolute performance level of the different reference workers in period 2.

formance in round 1 of the choosing participant, where D1 (D10) represents the 10 percent of the workers with the lowest (highest) performance in round 1. The figure reveals two important findings: First, irrespective of their own productivity in the first round, there is always a substantial proportion of participants who choose *not* to compare to any reference worker in the second round. While this share is largest among the lowest deciles (approximately 40 percent in D1-D3), there is also a significant share of the more productive workers that make a similar choice (about 25-30 percent in D7-D10).

Second, amongst participants who choose to compare to a reference worker, we find that most participants choose a reference worker whose performance is similar or higher to their own performance. The least productive participants (D1) predominantly choose to compare to the least productive reference worker (LO). The rest of the participants in the lower half of the productivity distribution (D2-D5) most frequently choose the intermediate reference worker (MI), while participants in the upper half of the distribution (D6-D10) mostly choose to compare to the best performing reference worker (HI).

Figure 6: Distribution of chosen reference workers (by productivity in round 1)



Note: The horizontal axis indicates the 10 different productivity deciles in the first round, ranked from the lowest productivity workers (D1) to the highest productivity workers (D10). Colors represent choice frequencies within a decile: Blue indicates the proportion of workers who choose *no* reference worker (NO), Red indicates the proportion of workers who choose the least productive reference worker (LO), Green indicates the proportion of workers who choose the average reference worker (MI), and Orange indicates the proportion of workers who choose the most productive reference worker (HI).

In Appendix C, we display the distributions of workers' productivity in round 1,

conditional on their chosen reference worker. These distributions clearly indicate that more productive participants tend to compare to more productive reference workers: The average productivity in round 1 of workers who pick the most productive reference worker is 1184, while it is 978 for those who pick the average reference worker and only 761 for those who pick least productive reference worker. In contrast, the average productivity in round 1 of those who choose not to compare to a reference worker is 997. In addition, we also show that choice patterns do not differ by gender (see Figure C2). Overall, these results indicate that a substantial part of the variation in workers' preferences for social comparisons can be explained by their productivity in the first round.

4.3.2 Uncovering workers' choice motives using text analysis

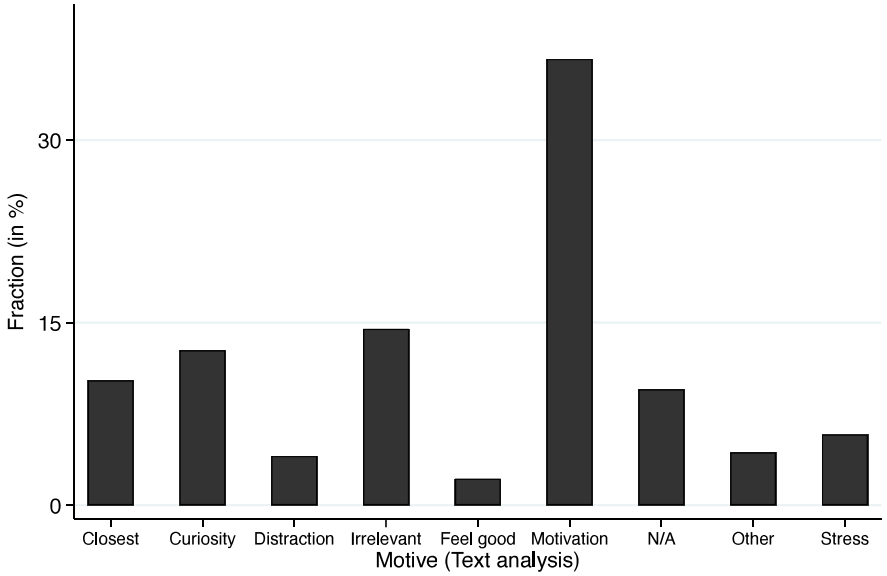
What are the main reasons invoked by workers to motivate their choices? Participants who were given the possibility to choose whom to compare to were asked to explain their decision in an open-text format. To unveil workers' motives and concerns, we hired a set of independent raters to code participants' answers. Raters were given a list of nine possible choice motives (which were identified through focus groups). Each rater was then asked to assign up to 3 different motives per worker. For example, the answer "*I chose to compare to this reference worker because it was the closest to me and I thought it would motivate me.*" could be assigned both to the category "Motivation" and to the category "Closest to me." We then aggregate raters' assessments at the worker-level by extracting the modal motive, i.e. the motive that is most often identified across raters.[33]

Figure 7 depicts the distribution of choice motives across all participants in ENDO (both subjects who choose to compare to one of the three possible reference worker and those who choose not to compare to a reference worker are included).

For 36.65 percent of the workers, "Motivation" is identified as the key determinant of their choice. These workers typically explain that they chose the option that they thought would help them be the most productive in round 2. For example, one such worker writes "*I wanted to compare myself to someone who had been faster than me so I could try to improve and keep up with them. If I had compared myself to someone slower, I wouldn't have been as motivated to go faster. If I didn't compare myself to anyone, I wouldn't have had that extra motivation to go even faster.*" 14.49 percent mention that choosing a reference worker was irrelevant for their performance and did not see any reason to compare with someone ("Irrelevant"). For example, one such worker writes "*I did not*

[33]We describe the details of the procedure for the text analysis in the Appendix D.

Figure 7: Distribution of choice motives



Note: The graph depicts the distribution of choice motives in ENDO. Each worker given the possibility to choose their reference worker was asked to explain their choice in an open-text format. Independent raters were asked to code participants' answers. Raters' assessments are then aggregated at the worker-level by extracting the modal motive (the motive that is the most often identified across raters).

think observing anyone would change the outcome of my performance." 10.25 percent report that the they chose the reference worker that was "Closest to them". For example, one worker explains "*I chose to observe the reference worker ranked 27 because their score was closest to mine. I didn't want to observe someone much lower or much higher because I knew they wouldn't help me be competitive. If I had picked the lowest one, then I would have no doubt I could win, and there would be no sense of competition, but if I picked the highest one, then I knew that goal was unreachable.*" 12.73 percent indicate that the choice was made out of "Curiosity" (e.g. "*to see how I stacked up against someone rated better.*") and 5.8 percent directly refer to "Stress" as a key driver of their choice (e.g. "*I didn't want the stress of watching someone else and trying to keep up*"). Last, a minority of 2.17 percent indicate that they chose whatever made them "feel good about themselves."[34]

Table 1 reveals how these motives relate to workers' choices. We document the distribution of choices (columns) as a function of the different motives (rows). For each motive, we highlight workers' modal choice in bold. Among workers who de-

[34]This diversity in choice motives also highlights potential heterogeneity in how stressful reference workers are perceived by subjects. While it is possible that some subjects perceive stress as not being particularly negative, the strong positive correlation between stress and self-reported nervousness suggests that this is not the case for most participants (discussed in footnote 30). Of course, whether and how negative these comparisons are perceived relates to whom people chose to compare to, as documented in Table 1). Moreover, note that subjects could always indicate not being stressed at all if comparisons did not bother them.

clare that their choice was mainly driven by a desire to motivate themselves, 79.94% picked the most productive reference worker while a minority of 14.12% (3.95%) chose to compare to the average (low) productivity reference worker and only 1.98% preferred not to compare to anyone. In contrast, workers who mentioned a desire to compare to someone close to themselves had a tendency to chose the intermediate reference worker (66.67%) while those who wanted to "feel good about themselves" predominantly picked the least productive reference worker (47.62%) or the intermediate reference worker (33.33%). Unsurprisingly, workers who i) said that comparing with someone else was irrelevant, ii) worried about their stress levels or iii) were concerned about being distracted mainly chose *not* to compare to a reference worker. Finally, curiosity leads a small portion of workers to predominantly compare with the most productive reference worker.

Table 1: Distribution of chosen reference workers (by choice motive)

	Chosen Reference Worker				
	NO	LO	MI	HI	Total
Motivation	1.98	3.95	14.12	**79.94**	100%
Closest to me	0	13.13	**66.67**	20.20	100%
Feel good about self	14.29	**47.62**	33.33	4.76	100%
Irrelevant	**99.29**	0	0	0.71	100%
Stress	**92.86**	3.57	1.79	1.79	100%
Distraction	**100**	0	0	0	100%
Curiosity	0	13.82	13.82	**72.36**	100%

Note: The table depicts the distribution of chosen reference workers (columns) as a function of the choice motive assigned to the worker by the independent raters (rows). For each motive (row), the modal choice is highlighted in bold. Each row sums up to 100 percent.

Taken together, Figure 6 and Table 1 illustrate that there is a wide variety of motives governing participants' choices in ENDO and that these motives result in a choice pattern that is substantially different from the matching pattern in EXBE. In particular, whereas almost all participants (91%) are matched with the most productive reference worker in EXBE, only 43% of the participants in ENDO choose to compare to this reference worker. This shift in the matching pattern is well aligned with the observation that stress increases much less strongly in ENDO than in EXBE. However, this raises the question why the lower frequency of matches with the best performing reference worker does not substantially impair the performance of participants in ENDO.

There are two plausible reasons that could explain why performance in ENDO and EXBE are virtually indistinguishable. First, while the right to choose whom to

compare to does, by definition, allow participants to be matched with their preferred reference worker, exogenous assignment—like in EXRA and EXBE—irremediably leads some workers to be "mismatched" with somebody they never wanted to observe. If exogenously assigned reference workers do not correspond to preferences, participants might be upset or disappointed. These mismatches might therefore substantially reduce the motivational spillovers of social comparisons. If this mismatch effect is important, the participants who did not pick the best performing reference worker in ENDO might not have improved their performance much if they had been forced to do so. They might however have perceived this experience as quite stressful.

Second, it is also possible that workers are a lot more affected by—or pay more attention to—a reference worker that they have chosen themselves as compared to the case where the same reference worker was exogenously assigned to them. Such a "choice effect"—if large enough—might (partly) compensate the decrease in motivation resulting from the fact that particpants in ENDO do often not choose the best performing reference worker.

In the following, we explore the relative importance of these two channels.

4.3.3 Disentangling the mismatch effect from the choice effect

To understand the difference between endogenously chosen and exogenously assigned social comparisons, it is not sufficient to simply compare the impact of the different reference workers in EXRA and in ENDO. The reason is that the participant populations who observe a particular reference worker in ENDO are self-selected, while they are randomly assigned in EXRA. We circumvent this issue by conducting a set of simulations that allow us to explore the relative importance of the mismatch effect and the choice effect.

We first use the data from EXRA to *simulate* what the average treatment effect would be if workers were exogenously assigned to the different reference workers in proportions that match the empirical distribution documented in ENDO.[35] Specifically, for each decile we randomly draw (without replacement) 100 subjects from the different treatments arms of EXRA in proportions that match the distribution of chosen reference workers in that *same* decile in ENDO (i.e. according to Figure 6). For example, out of the 100 workers in ENDO who were the *least* productive in round 1 (D1), 40 selected no reference worker, 30 selected LO, 10 selected MI and 20 selected HI. We *reconstruct* this sample by randomly drawing, out of the least productive workers in EXRA (D1), 40 who were assigned to EXRA-NO, 30 who were assigned to

[35] For expositional clarity, this example only refers to participants' productivity in period 1. However, our simulations also take participants' gender into account.

EXRA-LO, 10 who were assigned to EXRA-MI and 20 who where assigned to EXRA-HI. We do the same for the remaining nine deciles and then calculate the average treatment effect for this *simulated* sample. We repeat this procedure 1000 times (i.e. we calculate the average treatment effect for each of the 1000 simulated samples), and then compute the average over these 1000 simulated treatments effects.

This simulated treatment effect ("Simulation 1") informs us about the effect of exogenously assigned reference workers when they are assigned in proportion that match the empirical distribution in ENDO. While this simulated sample shares the main characteristics of the ENDO sample with respect to workers' observable characteristics (their productivity in round 1 and their gender), it does *not* take workers' intrinsic preferences over reference workers into account. This simulation therefore most likely includes many participants who were not assigned to their preferred reference worker, i.e. it includes "mismatched" participants.

To identify the size of the mismatch effect, we repeat the simulation exercise described above by *exclusively* drawing (again without replacement) individuals from the different EXRA treatment arms who were assigned to the reference worker they would have chosen anyways, i.e. we systematically account for workers' preferences for peers.[36] For example, we only draw workers from EXRA-HI who revealed to us that their preferred reference worker would have been HI. This procedure allows us to reconstruct a second simulated sample ('Simulation 2') which shares the main characteristics of the ENDO sample *both* with respect to workers' observable characteristics (their productivity in round 1 and their gender) *and* their preferences over reference workers.

These two simulated samples allow us to determine the relative importance of the mismatch effect and the choice effect. The mismatch effect is captured by the difference between the two simulated samples. If participants' preferences over reference workers matter so that exogenously created mismatches *reduce* the performance enhancing effect of social comparisons, then we should find that the average performance in the second simulation *exceeds* the one in the first simulation (Simulation 2 – Simulation 1 > 0). In contrast, the absence of such an effect (i.e. if the difference between the two simulations is negligible) would imply that what matters is which reference worker individuals are assigned to, but not whether this reference worker coincides with their preferred alternative.

The choice effect is captured by the difference between ENDO and the second

[36] For each participant in EXRA, we elicited information about that individual's *preferred* reference worker, i.e. we asked each worker to tell us which of the three potential reference worker—if any—they would have liked to compare themselves to if they had the possibility to choose. Participants could also indicate if they would have preferred not to compare to anyone.

simulated sample. If participants care more about endogenously chosen reference workers than about exogenously assigned ones (even though they are matched to their preferred reference worker in both cases), we should find that the productivity gains in ENDO exceed those in Simulation 2. However, if the right to choose per se has no effect (i.e. if there is no choice effect), then we should find no such difference.

We depict the counterfactual treatment effects for these two simulated samples, along with those in EXRA and ENDO, in the Figure 8 below. The figure reveals two striking findings.

First, the counterfactual increase in performance in Simulation 2 is larger than in Simulation 1 (Simulation 1 : + 126.5 units, Simulation 2 : + 137.75 units). This result suggests that there is a substantial mismatch effect, i.e. that exogenously assigned reference workers are indeed *less* effective in motivating participants if they do *not* align with participants' preferences. As a consequence, a policy aimed at maximizing motivational spillovers should, as much as possible, take these preferences into account in order to succeed.

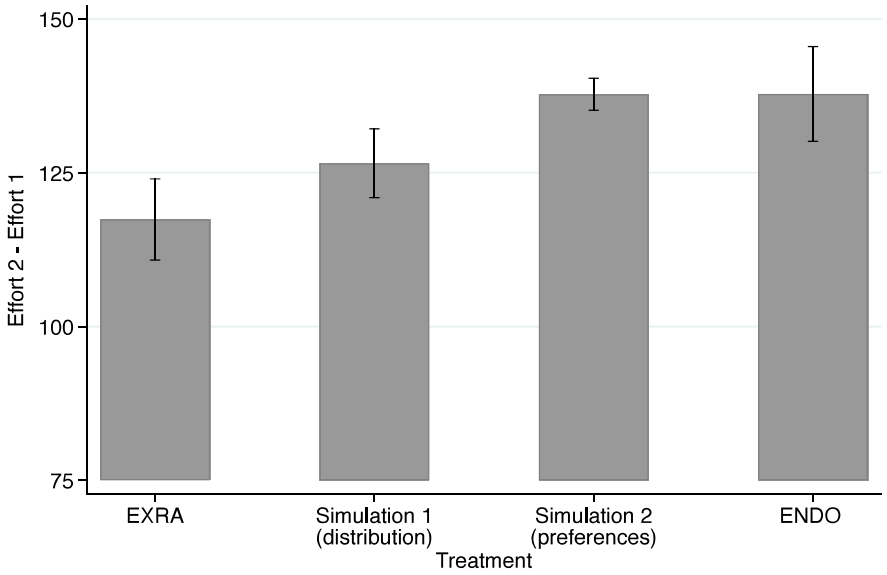
Second, the counterfactual increase in performance in Simulation 2 matches almost perfectly the one observed in ENDO (Simulation 2 : + 137.75 units, ENDO : + 138 units). This finding indicates that the possibility to choose does *not*, in itself, affect participants as long as they are matched with their preferred reference worker, i.e. there is no choice effect.

Overall, these results underscore the importance of respecting workers' preference when matching them with a reference worker. While endogenously chosen social comparisons do, by definition, allow workers to observe their preferred peer; forcing them to compare to other reference workers generates mismatches that are ultimately detrimental to the overall performance of the workforce. These findings explain why a targeted exogenous assignment procedure aimed at maximizing motivational spillovers (EXBE) does not substantially outperform a matching system with endogenous choice (ENDO). Moreover, targeted matching increases perceived stress substantially, because the predicted most motivating reference workers are also the ones that are perceived as most stressful to observe. This higher stress level is unnecessary for those workers who do not desire to be matched with the best performing reference worker, because the additional motivational effect from forcing these worker to observe a top performing peer tend to be very limited.

Summarizing, our data reveal that endogenous choice of peers is very effective, because it enables those workers who are interested in getting a motivational boost to pick a highly motivating peer. At the same time, it prevents those workers who prefer a different reference worker or who do not want to be matched to anyone from

experiencing unnecessary high levels of stress.

Figure 8: Understanding the differences between exogenously assigned and endogenously chosen reference workers: mismatch effects vs. choice effect.



Note: EXRA and ENDO document the average treatment effect in these two treatments, respectively. 'Simulation 1' and 'Simulation 2' correspond to the two counterfactual, simulated, treatment effects (based on 1000 iterations). 'Simulation 1' depicts the simulated treatment effect in a counterfactual treatment in which workers are exogenously assigned to reference workers in proportions that match the empirical distribution of participants' choices in ENDO. 'Simulation 2' replicates 'Simulation 1', with the additional constraint to only draw workers from the different EXRA treatment arms who where exogenously assigned to their preferred reference worker (if any). Whiskers represent the standard errors.

4.4 Benchmarking the effects of social comparisons and robustness

4.4.1 Monetary incentives

So far, our analysis has highlighted the role of social comparisons and different assignment mechanisms for productivity and for stress. Important questions that were left unanswered up to this point are whether these effects are economically meaningful and robust. One might be concerned, for example, that the magnitude of these effects is small in comparison to the productivity gains that can be achieved using standard economic tools such as performance pay. One might also worry that the impact of social comparisons vanishes in the presence of financial incentives.

To adress these possible concerns, we conducted three additional treatments (RANK×$, ENDO×$, EXBE×$) in which social comparisons are combined with financial incentives for production. These treatments are exactly identical to the original

treatments described above (RANK, ENDO, EXBE), with the exception that workers are unexpectedly offered a piece rate of 1 cent per 100 units of output produced in period 2 in addition to their fixed payment.[37]

We report the effects of these treatments in Appendix B.4. Three important insights emerge from this analysis. First, we find that participants' response to financial incentives is in line with the predictions of standard economic theory. RANK×$ generates an increase in performance that is more than *twice* the size of the increase in performance in RANK ($p < 0.01$). On average, the increase in productivity of workers which are paid a piece-rate (i.e. pooling RANK×$, ENDO×$, EXBE×$) is 53% larger than the increase in productivity of the workers in the equivalent treatments without the piece rate (i.e. pooling RANK, ENDO and EXBE; $p < 0.01$). While these financial incentives have positive effects on performance, they also generates a significant additional increase in stress of about 13 percent ($p < 0.01$).[38]

Second, we show that social comparisons alone can generate productivity gains that are of the same magnitude as those achieved through the introduction of a piece rate. Indeed, the average increase in performance in RANK×$ is statistically indistinguishable from the effects observed in EXBE and in ENDO (RANK×$ vs. EXBE : $p = 0.94$; RANK×$ vs. ENDO : $p = 0.62$). Interestingly, the increase in stress documented in RANK×$ (+0.58 points, approximately + 45 percent of a standard deviation, $p < 0.01$) is not significantly different from the one observed in ENDO (test of difference, $p = 0.18$), but is significantly *lower* than the one reported in EXBE (test of difference, $p < 0.01$). These results suggest that social incentives can be a very effective and cheap way of motivating the workforce, and that letting workers choose whom to compare to can generate economically meaningful behavioral effects without causing an excessive increase in stress amongst the workers.

Third, all the main empirical regularities that we documented throughout the paper are robust to the introduction of steeper financial incentives: i) financial incentives do virtually *not* affect whom workers choose to compare to (see Figure B.2 in Appendix B.4), and ii) financial incentives do *not* wipe out the effects of social comparisons, i.e. social comparisons still boost productivity even when interacted with monetary rewards. Just like in the treatments without the piece rate, letting workers choose whom to compare to (ENDO×$) generates an increase in productivity that is of roughly the same magnitude as when forcing them to compare to the most moti-

[37]This amounts to an average additional 10-15 cents, which is a substantial pay increase for a 5 minutes task on MTurk as it corresponds to an approximate 10% increase in pay (see DellaVigna and Pope, 2017, for a discussion).

[38]The average increase in stress in the treatments with financial incentives is of +0.68 points (+53 percent of a standard deviation), whereas it is of +0.6 points (+47 percent of a standard deviation) in the corresponding treatments that do not include financial incentives.

vating reference worker (EXBE×$)[39] but it also yields a substantially smaller increase in perceived stress ($p < 0.05$). Overall, these results suggest that our findings are not driven by the specificities of the monetary rewards used to incentivize the workforce, and that social incentives and monetary rewards act as complements in our setting.

4.4.2 Non-social comparisons

Our study focuses on motivational spillovers stemming from *social* comparisons. However, one might wonder how important it is for our results that the comparisons are indeed social. In particular, could it be that participants would react in the exact same way to comparable but *non-social* reference points? While it seems plausible that our subjects interpret the performance of their reference worker as a goal to attain, it remains an open question whether exposing them to non-social information would generate similar effects.[40] We answer this question by conducting an additional pre-registered experiment, in which we compare the participants who observe a highly productive reference worker with participants who are confronted with an equally challenging non-social pacemaker.[41]

In this additional study we randomly allocate 500 participants to the EXRA-HI treatment, while another 500 participants are randomly assigned to a "pacemaker" condition (PACE-HI)—a non-social version of the EXRA-HI treatment. Like the real-time performance of the reference workers in our social treatments, the pacemaker is also displayed as a growing vertical bar (whose constant speed is set to reach exactly the same number of points as the reference worker in the EXRA-HI treatment). The two treatments are therefore identical except for the fact that in EXRA-HI the increasing bar represents the performance of another human being, while in PACE-HI the pacemaker does not provide any information about the performance of peers. While neither treatment explicitly mentions that participants are expected to keep up with the growing vertical bar, we purposefully use the word "pacemaker" (which arguably has a goal-related connotation) in the PACE-HI condition.

If both the performance of the reference worker and the pacemaker are interpreted as goals to attain, then we would expect PACE-HI and EXRA-HI to generate a comparable increase in performance. Alternatively, if social comparisons lead to motivational spillovers that go beyond the effects of arbitrarily set non-social goals,

[39] Although note that the difference is marginally significant ($p = 0.09$).

[40] It is also possible that some participants set their own internal goal. However, this is true across all treatments and can therefore not explain the treatment effects reported throughout the paper, suggesting that social and non-social comparisons replace self-set goals (if any).

[41] We preregistered this additional study as trial number 137539 on AsPredicted.org. For details on the design, see Appendix B.6.

then we should observe a larger performance increase in EXRA-HI.

Our results unambiguously show that motivational effects triggered by social comparisons surpass those brought about by otherwise identical non-social goals. As a matter of fact, the increase in performance in EXRA-HI is more than twice as large as the one in PACE-HI (test of difference: $p < 0.01$, see Table B.12 in Appendix B.6).

Social comparisons not only have larger effects on workers' performance, they also have very different effects on workers' perceptions (see Tables B.12 and B.13 in Appendix B.6). Indeed, workers in the EXRA-HI condition report being substantially more stressed ($p < 0.01$) and more nervous ($p < 0.05$) than workers in the pacemaker condition. They are also more likely to report that the comparison i) motivated them ($p < 0.01$), ii) generated a greater feeling of competition ($p < 0.05$), and iii) positively affected their performance ($p < 0.01$) than participants assigned to the non-social pacemaker condition.

Altogether, these results indicate that the social aspect of output comparisons is a key driver of our findings, i.e. social comparisons generate much larger behavioral effects than comparable non-social goals.

5 Conclusions

Individuals frequently compare themselves with others, and the workplace is no exception. While the existing literature has primarily focused on the role of randomly assigned peers for productivity, existing work has largely neglected the question whether alternative peer assignment procedures can enhance productivity by leveraging social comparisons. We study this question using a large scale real-effort experiment. Moreover, whereas researchers and practitioners previously focused on the effects of social comparisons for primary outcome variables (e.g. productivity), we show that they also have important effects for workers' perceived stress.

Our results reveal that peer assignment mechanisms importantly shape the behavioral effects of social comparisons. Workers who are exogenously assigned to their predicted most motivating peer and those who endogenously choose whom to compare to are both significantly more productive than those assigned to a random peer. However, endogenous choice generates a much smaller increase in perceived stress than exogenous assignment to the predicted most motivating peer. We show that the desirable effects of endogenous peer choice can be explained by two factors: First, those workers whose peer choice is guided by the desire to enhance their performance predominantly choose to compare with a highly productive peer. This pro-

portion of the workforce is large enough to create a substantial performance increase. Second, workers who perceive social comparisons to be stressful, distracting or irrelevant do mostly not compare themselves with a reference worker. This "opt-out" strategy allows them to keep their stress level relatively low. They also do not lose much in terms of motivational spillovers because forcing them to compare to a peer they would not have chosen would only have had a small effect on their performance due to the mismatch effect. Last, we show that social comparisons generate much larger motivational effects than comparable non-social goals—and that the magnitude of these effects is similar to those of a monetary incentive that increases pay by about 10 percent.

Collectively, our results suggest that social comparisons can in principle be leveraged to boost productivity, but they also highlight that different policies can have different (negative) unintended consequences such as, for example, raising the perceived stress of the workers. Although outside of the scope of this paper, these results suggest that the welfare implications of different policies can be debated—consistent with the recent discussion on the welfare effects of "social nudges" (Allcott and Kessler, 2019; Butera et al., 2022). Thus, we believe that an important implication of our paper is that the "plausible, unintended consequences" of policies should be measured more systematically. For example, while a large literature in (personnel) economics has focused on how to best incentivize the workforce (e.g. by comparing the effects of fixed payment versus tournament incentives), it has often neglected to evaluate the impacts of such policies for important outcomes such as workers' stress, satisfaction, or psychological well-being. However, these dimensions should also matter for policy makers and practitioners since these are also facets of workers' experience that ultimately affect the firms' performance. Whether and how companies should trade-off these dimensions is a particularly exciting open question for future research.

Our experimental design was purposefully kept relatively stylized in order to cleanly identify the effects of social comparisons. For example, production complementarities as well as learning spillovers were excluded by design, the task was short lived and unlikely to convey ego-relevant information, and workers remained anonymous. Whether and how these elements interact with social comparisons remains an open question which is beyond the scope of our paper, which was to cleanly isolate the motivational effects of different peer assignment mechanisms in the simplest environment possible. We see our study as an important first step towards understanding motivational spillovers, and believe that future studies can easily extend our design to explore these exciting questions.

Moreover, while our study is concerned with the effects of social comparisons in

the context of work on an effort intensive task, social comparisons have been shown to matter in many other settings as well (e.g., education, voting, or energy consumption, among others). In these settings, too, it might be possible to leverage social spillovers using non-random assignment mechanisms (for preliminary evidence on the effects of non-random peers in the education context, see, for example, Carrell et al., 2013; Chen and Gong, 2018; Fischer et al., 2021; Kiessling et al., 2021). Our methodology could easily be adapted and extended to these contexts.

Finally, our results and methodology might also be useful to social scientists interested in the nature of social comparisons more generally. Social comparisons have been studied for a long time (see, e.g., Festinger 1954; Frank 1985) and they play a central role in many recent theoretical developments, ranging from models of inequity aversion (see e.g. Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000; Charness and Rabin, 2002), to theories of conformism (Bernheim, 1994) and social image (see e.g. Bénabou and Tirole, 2006), among others. While these models typically take the relevant social reference group as exogenously given, empirical evidence on whom people actually compare themselves to—and on the determinants of these choices—remains very scarce. We hope that our paper will spark new research in this important area as well.

References

Allcott, Hunt and Judd B Kessler, "The welfare effects of nudges: A case study of energy use social comparisons," *American Economic Journal: Applied Economics*, 2019, *11* (1), 236–76.

Almås, Ingvild, Alexander W Cappelen, and Bertil Tungodden, "Cutthroat capitalism versus cuddly socialism: Are Americans more meritocratic and efficiency-seeking than Scandinavians?," *Journal of Political Economy*, 2020, *128* (5), 1753–1788.

Amir, On and Dan Ariely, "Resting on laurels: The effects of discrete progress markers as subgoals on task performance and preferences.," *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 2008, *34* (5), 1158.

Ashraf, Nava and Oriana Bandiera, "Social incentives in organizations," *Annual Review of Economics*, 2018, *10*, 439–463.

Bandiera, Oriana, Iwan Barankay, and Imran Rasul, "Social preferences and the response to incentives: Evidence from personnel data," *The Quarterly Journal of Economics*, 2005, pp. 917–962.

_ , _ , and _ , "Social incentives in the workplace," *The Review of Economic Studies*, 2010, *77* (2), 417–458.

Banerjee, Abhijit, Esther Duflo, Amy Finkelstein, Lawrence F Katz, Benjamin A Olken, and Anja Sautmann, "In praise of moderation: Suggestions for the scope and use of pre-analysis plans for rcts in economics," Technical Report, National Bureau of Economic Research 2020.

Bárcena-Martín, Elena, Alexandra Cortés-Aguilar, and Ana I Moro-Egido, "Social comparisons on subjective well-being: The role of social and cultural capital," *Journal of Happiness Studies*, 2017, *18* (4), 1121–1145.

Bénabou, Roland and Jean Tirole, "Incentives and prosocial behavior," *American economic review*, 2006, *96* (5), 1652–1678.

Berger, Jonah and Devin Pope, "Can losing lead to winning?," *Management Science*, 2011, *57* (5), 817–827.

Bernheim, B Douglas, "A theory of conformity," *Journal of Political Economy*, 1994, *102* (5), 841–877.

Beugnot, Julie, Bernard Fortin, Guy Lacroix, and Marie Claire Villeval, "Gender and peer effects on performance in social networks," *European Economic Review*, 2019, *113*, 207–224.

Bó, Pedro Dal, Andrew Foster, and Louis Putterman, "Institutions and behavior: Experimental evidence on the effects of democracy," *American Economic Review*, 2010, *100* (5), 2205–29.

Bolton, Gary E and Axel Ockenfels, "ERC: A theory of equity, reciprocity, and competition," *American Economic Review*, 2000, *90* (1), 166–193.

Bursztyn, Leonardo and Robert Jensen, "How does peer pressure affect educational investments?," *The Quarterly Journal of Economics*, 2015, *130* (3), 1329–1367.

_ **and** _ , "Social image and economic behavior in the field: Identifying, understanding, and shaping social pressure," *Annual Review of Economics*, 2017, *9*, 131–153.

_ **, Florian Ederer, Bruno Ferman, and Noam Yuchtman**, "Understanding mechanisms underlying peer effects: Evidence from a field experiment on financial decisions," *Econometrica*, 2014, *82* (4), 1273–1301.

_ **, Georgy Egorov, and Robert Jensen**, "Cool to be smart or smart to be cool? Understanding peer pressure in education," *The Review of Economic Studies*, 2019, *86* (4), 1487–1526.

Butera, Luigi, Robert Metcalfe, William Morrison, and Dmitry Taubinsky, "Measuring the welfare effects of shame and pride," *American Economic Review*, 2022, *112* (1), 122–68.

Buunk, Abraham P and Pieternel Dijkstra, "Social comparisons and well-being," in "The happy mind: Cognitive contributions to well-being," Springer, 2017, pp. 311–330.

Cappelen, Alexander W, Cornelius Cappelen, and Bertil Tungodden, "Second-Best Fairness: The Trade-off between False Positives and False Negatives," *American Economic Review*, 2021.

_ **, Karl Ove Moene, Siv-Elisabeth Skjelbred, and Bertil Tungodden**, "The merit primacy effect," *The Economic Journal*, 2023, *133* (651), 951–970.

Card, David, Alexandre Mas, Enrico Moretti, and Emmanuel Saez, "Inequality at work: The effect of peer salaries on job satisfaction," *American Economic Review*, 2012, *102* (6), 2981–3003.

Carrell, Scott E, Bruce I Sacerdote, and James E West, "From natural variation to optimal policy? The importance of endogenous peer group formation," *Econometrica*, 2013, *81* (3), 855–882.

Charness, Gary and Matthew Rabin, "Understanding social preferences with simple tests," *The Quarterly Journal of Economics*, 2002, *117* (3), 817–869.

_ , David Masclet, and Marie Claire Villeval, "The dark side of competition for status," *Management Science*, 2013, *60* (1), 38–55.

Chen, Roy and Jie Gong, "Can self selection create high-performing teams?," *Journal of Economic Behavior & Organization*, 2018, *148*, 20–33.

Chen, Yan, F Maxwell Harper, Joseph Konstan, and Sherry Xin Li, "Social comparisons and contributions to online communities: A field experiment on movielens," *American Economic Review*, 2010, *100* (4), 1358–98.

Cicala, Steve, Roland G Fryer, and Jörg L Spenkuch, "Self-selection and comparative advantage in social interactions," *Journal of the European Economic Association*, 2018, *16* (4), 983–1020.

Clark, Andrew E and Claudia Senik, "Who compares to whom? The anatomy of income comparisons in Europe," *The Economic Journal*, 2010, *120* (544), 573–594.

Coffman, Lucas C, Clayton R Featherstone, and Judd B Kessler, "Can social information affect what job you choose and keep?," *American Economic Journal: Applied Economics*, 2017, *9* (1), 96–117.

Corgnet, Brice, Joaquín Gómez-Miñambres, and Roberto Hernán-Gonzalez, "Goal setting and monetary incentives: When large stakes are not enough," *Management Science*, 2015, *61* (12), 2926–2944.

_ , _ , and _ , "Goal setting in the principal–agent model: Weak incentives for strong performance," *Games and Economic Behavior*, 2018, *109*, 311–326.

Cullen, Zoë and Ricardo Perez-Truglia, "How much does your boss make? The effects of salary comparisons," *Journal of Political Economy*, 2022, *130* (3), 766–822.

DellaVigna, Stefano and Devin Pope, "What motivates effort? Evidence and expert forecasts," *The Review of Economic Studies*, 2017, *85* (2), 1029–1069.

_ , John A List, Ulrike Malmendier, and Gautam Rao, "Voting to tell others," *The Review of Economic Studies*, 2016, *84* (1), 143–181.

Esopo, Kristina, Johannes Haushofer, Linda Kleppin, and Ingvild Skarpeid, "Acute stress decreases competitiveness among men," Technical Report, Working paper 2019.

Falk, Armin and Andrea Ichino, "Clean evidence on peer effects," *Journal of Labor Economics*, 2006, *24* (1), 39–57.

_ **and Markus Knell**, "Choosing the Joneses: Endogenous goals and reference standards," *Scandinavian Journal of Economics*, 2004, *106* (3), 417–435.

Fehr, Ernst and Klaus M Schmidt, "A theory of fairness, competition, and cooperation," *The Quarterly Journal of Economics*, 1999, *114* (3), 817–868.

Festinger, Leon, "A theory of social comparison processes," *Human Relations*, 1954, *7* (2), 117–140.

Fischer, Mira, Rainer Michael Rilke, and B Burcin Yurtoglu, "When, and Why, Do Teams Benefit from Self-Selection?," 2021.

Frank, Robert H, *Choosing the right pond: Human behavior and the quest for status.*, Oxford University Press, 1985.

Fujita, Frank and Ed Diener, "Social comparisons and subjective well-being," *Health, coping and well-being: Perspectives from social comparison theory*, 1997, pp. 329–357.

Gerber, Alan S, Donald P Green, and Christopher W Larimer, "Social pressure and voter turnout: Evidence from a large-scale field experiment," *American Political Science Review*, 2008, *102* (1), 33–48.

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

Graham, Bryan S, Guido W Imbens, and Geert Ridder, "Complementarity and aggregate implications of assortative matching: A nonparametric analysis," *Quantitative Economics*, 2014, *5* (1), 29–66.

Halkos, George and Dimitrios Bousinakis, "The effect of stress and satisfaction on productivity," *International Journal of Productivity and Performance Management*, 2010.

Haushofer, Johannes and Jeremy Shapiro, "The short-term impact of unconditional cash transfers to the poor: experimental evidence from Kenya," *The Quarterly Journal of Economics*, 2016, *131* (4), 1973–2042.

_ , **Chaning Jang, John Lynham, and Justin Abraham**, "Stress and temporal discounting: Do domains matter," *mimeo*, 2015.

_ , **Prachi Jain, Abednego Musau, and David Ndetei**, "Stress may increase choice of sooner outcomes, but not temporal discounting," *Journal of Economic Behavior & Organization*, 2021, *183*, 377–396.

Herbst, Daniel and Alexandre Mas, "Peer effects on worker output in the laboratory generalize to the field," *Science*, 2015, *350* (6260), 545–549.

Horton, John J, David G Rand, and Richard J Zeckhauser, "The online laboratory: Conducting experiments in a real labor market," *Experimental Economics*, 2011, *14* (3), 399–425.

Jacobson, Bert H, Steven G Aldana, Ron Z Goetzel, KD Vardell, Troy B Adams, and Rick J Pietras, "The relationship between perceived stress and self-reported illness-related absenteeism," *American Journal of Health Promotion*, 1996, *11* (1), 54–61.

Kiessling, Lukas, Jonas Radbruch, and Sebastian Schaube, "Self-selection of peers and performance," *Management Science*, 2021.

_ , _ , _ **et al.**, "Determinants of Peer Selection," Technical Report, University of Bonn and University of Mannheim, Germany 2019.

Kirchler, Michael, Florian Lindner, and Utz Weitzel, "Rankings and risk-taking in the finance industry," *The Journal of Finance*, 2018, *73* (5), 2271–2302.

Koivisto, Jonna and Juho Hamari, "The rise of motivational information systems: A review of gamification research," *International Journal of Information Management*, 2019, *45*, 191–210.

Kräkel, Matthias, "Peer effects and incentives," *Games and Economic Behavior*, 2016, *97*, 120–127.

Kube, Sebastian, Michel André Maréchal, and Clemens Puppe, "The currency of reciprocity: Gift exchange in the workplace," *American Economic Review*, 2012, *102* (4), 1644–62.

_ , _ , **and** _ , "Do wage cuts damage work morale? Evidence from a natural field experiment," *Journal of the European Economic Association*, 2013, *11* (4), 853–870.

Lazear, Edward P, "Performance Pay and Productivity," *The American Economic Review*, 2000, *90* (5), 1346–1361.

Leontaridi, Rannia M and Melanie E Ward-Warmedinger, "Work-related stress, quitting intentions and absenteeism," *Quitting Intentions and Absenteeism (May 2002)*, 2002.

Manski, Charles F, "Identification of endogenous social effects: The reflection problem," *The Review of Economic Studies*, 1993, *60* (3), 531–542.

Mas, Alexandre and Enrico Moretti, "Peers at Work," *American Economic Review*, 2009, *99* (1), 112–145.

Mosadeghrad, Ali Mohammad, "Occupational stress and turnover intention: implications for nursing management," *International Journal of Health Policy and Management*, 2013, *1* (2), 169.

Perez-Truglia, Ricardo, "The effects of income transparency on well-being: Evidence from a natural experiment," *American Economic Review*, 2020, *110* (4), 1019–1054.

Quidt, Jonathan De, Johannes Haushofer, and Christopher Roth, "Measuring and bounding experimenter demand," *American Economic Review*, 2018, *108* (11), 3266–3302.

Roels, Guillaume and Xuanming Su, "Optimal design of social comparison effects: Setting reference groups and reference points," *Management Science*, 2014, *60* (3), 606–627.

Sacerdote, Bruce, "Peer effects with random assignment: Results for Dartmouth roommates," *The Quarterly journal of economics*, 2001, *116* (2), 681–704.

Schwerter, Frederik, "Social Reference Points and Risk Taking," 2019.

Shearer, Bruce, "Piece rates, fixed wages and incentives: Evidence from a field experiment," *The Review of Economic Studies*, 2004, *71* (2), 513–534.

Snowberg, Erik and Leeat Yariv, "Testing the waters: Behavior across participant pools," *American Economic Review*, 2021, *111* (2), 687–719.

Suls, Jerry, Rene Martin, and Ladd Wheeler, "Social comparison: Why, with whom, and with what effect?," *Current Directions in Psychological Science*, 2002, *11* (5), 159–163.

Villeval, Marie Claire, "Performance Feedback and Peer Effects," Technical Report, GLO Discussion Paper 2020.

A Background information on experimental task and population sample

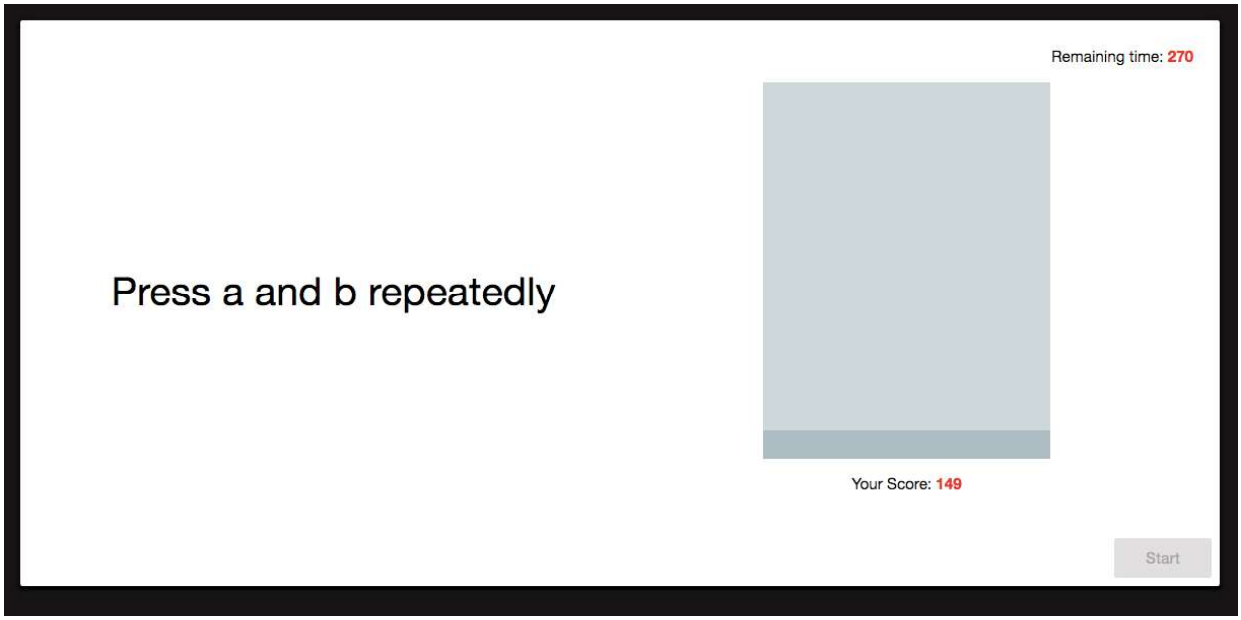
A.1 The experiment

A.1.1 The EXRA treatment

Figure A.1 is a screenshot of the real effort task in round 1, during which the workers only get information about their *own* production. On the screen, workers find a reminder of the instructions ("Press a and b repeatedly"). They are also informed about their current output (which is represented both numerically and graphically with a growing vertical bar) and about the remaining time to complete the assignment.

This screenshot also corresponds to what *some* of the workers see during the second production round. Indeed, workers in the RANK treatment, workers who are exogenously assigned to no reference worker (EXRA-NO), and workers in the ENDO treatment who choose not to compare to another worker complete the task in round 2 in the exact same conditions as in round 1, i.e. with information about their own production only. Note that it also corresponds to the screen seen by the workers from the reference population, since they had to complete both production rounds while only seeing information about their own production.

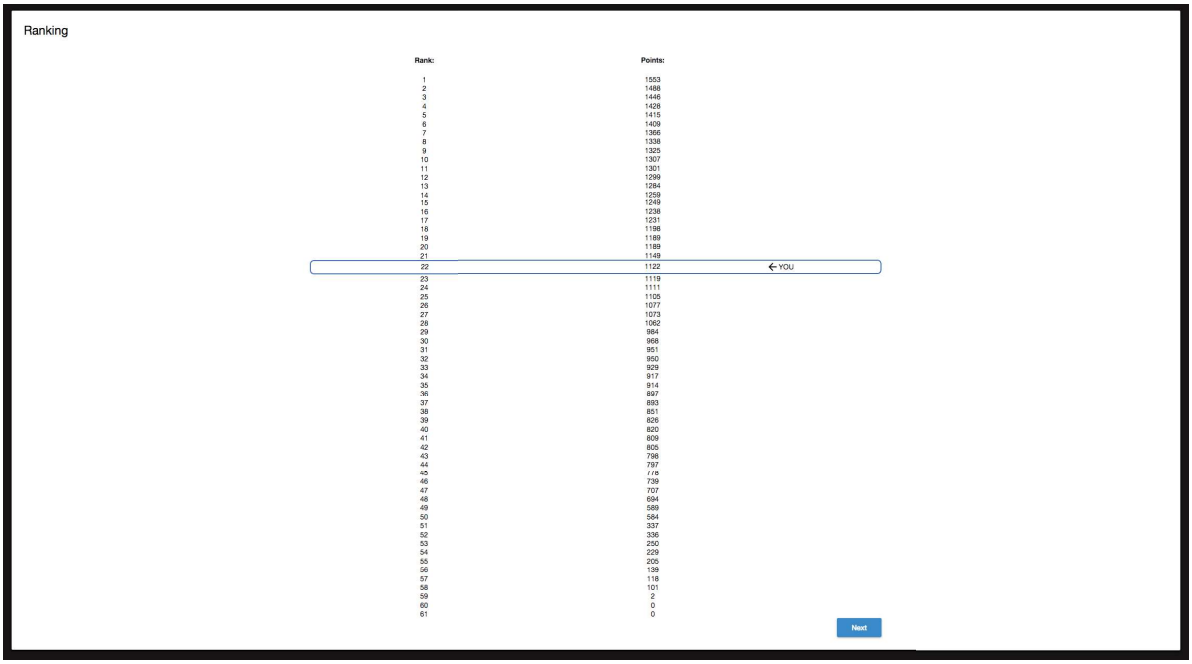
Figure A.1: Real effort task with information about own production only



Upon completion of round 1, participants in all the treatments are compared to

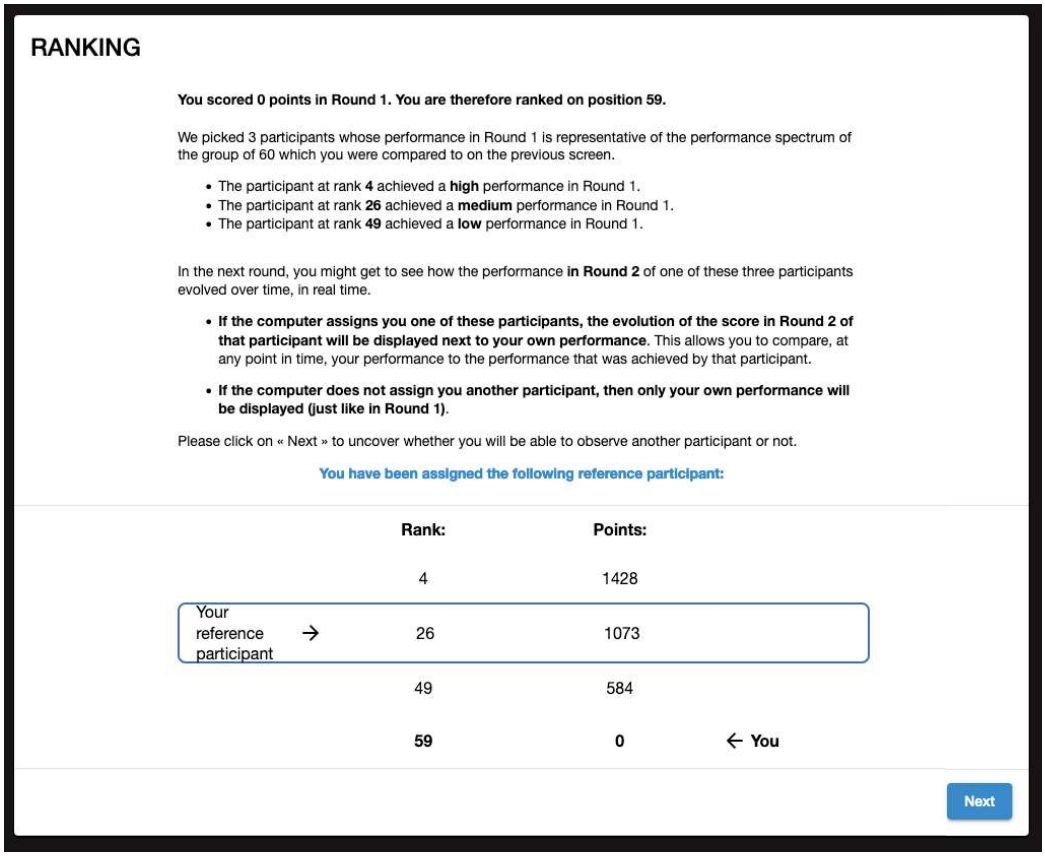
the reference population, i.e. we compare the performance in round 1 of our participants with the performance in round 1 of the workers from the reference population (See Figure A.2). This stage allows participants to inform themselves about their rank and their output, and to compare it with the rank and the output in round 1 of *all* the 60 workers from the reference population. The information related to the worker's own rank and production is highlighted in blue.

Figure A.2: Ranking stage.



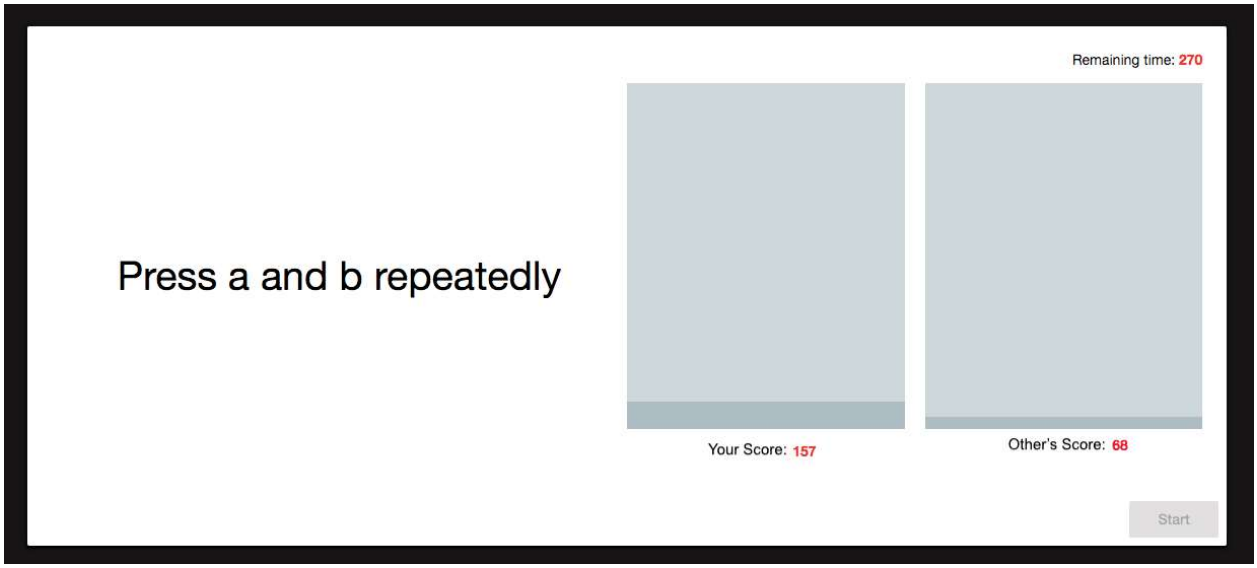
After seeing their rank, participants then learn that they will have to complete the task a second time. In the EXRA treatment, they are informed that they might have the possibility to compare themselves to another worker who completed the same task in the past (see Figure A.3), while working on the task a second time. They learn about the three possible reference workers that they might be assigned to, and about the consequences of being assigned to one of them (or to none) for the second production round.

Figure A.3: Exogenous assignment to one (or no) reference worker. In this example, the participant is assigned to the reference worker with average productivity (EXRA-MI, ranked 26).



In round 2, participants who are assigned to one of the three possible reference workers not only receive information about their own production in round 2, but they also receive real-time information about the production in round 2 of the reference worker they have been assigned. This information is depicted both numerically and graphically as a second growing vertical bar (See Figure A.4). Such a screen can be encountered in round 2 by the participants who see a reference worker, i.e. those who are exogenously assigned to a reference worker (EXRA-LO, EXRA-MID, EXRA-HI, EXBE) or those in the ENDO who choose to compare to a reference worker.

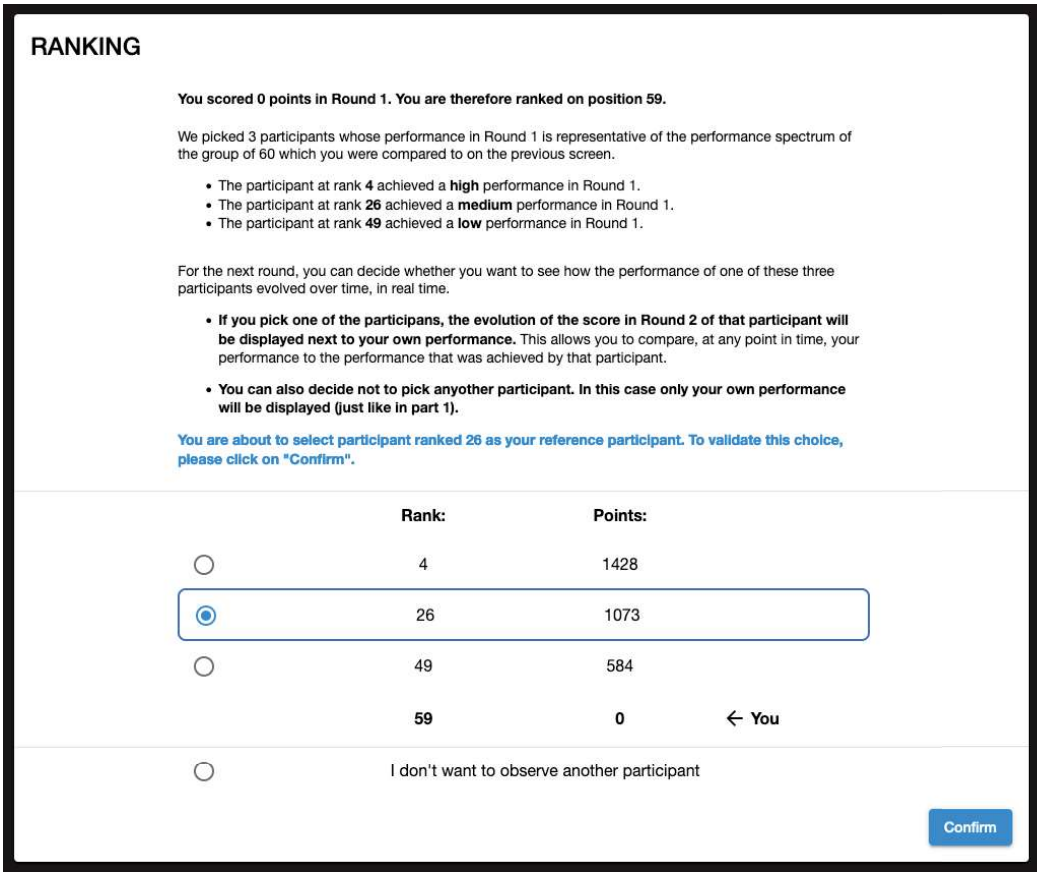
Figure A.4: Real effort task in round 2 during which the worker receives real-time information about the production of a reference worker.



A.1.2 The ENDO treatment

For participants in the ENDO treatment, the sequence of events and the screenshots are similar to those depicted above. The only difference is that, after the rank stage (Figure A.2), participants in ENDO are informed that they will get the possibility—if they would like to—to compare to another worker while they complete round 2 (see Figure A.5). They are informed about the potential reference workers they can choose from, and are asked to choose to whom they want to compare (if at all). They are also informed about the consequences of their choice for what will happen in the next production round. To keep things as comparable as possible with the EXRA and the EXBE treatment, the wording of the entire screen is kept identical.

Figure A.5: Choice of reference worker (ENDO treatment). This screen is similar to the one shown to participants in EXRA/EXBE in which the exogenous assignment procedure to reference workers is explained, with the exception participants in the ENDO treatment can decide whom to compare to.



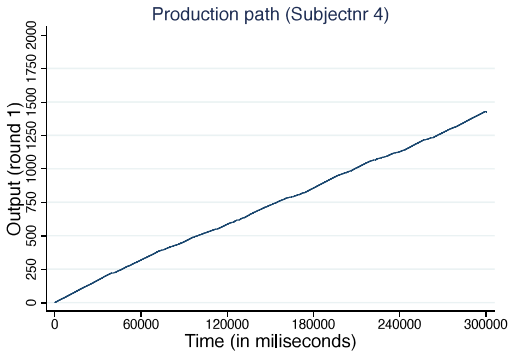
A.2 Reference population and potential reference workers

Table A.1 depicts the ranking (rank) of the 60 workers from the reference population, sorted by their output in round 1 (effort1). The table also depicts the baseline participants' identifying number (subjectnr) as well as their round 2 output (effort2). The three potential reference workers (rank 4, 26 and 49) are highlighted **bold**. Figures A.6 to A.8 depict the production path in round 1 (panel a) and 2 (panel b) for each of these 3 potential reference workers.

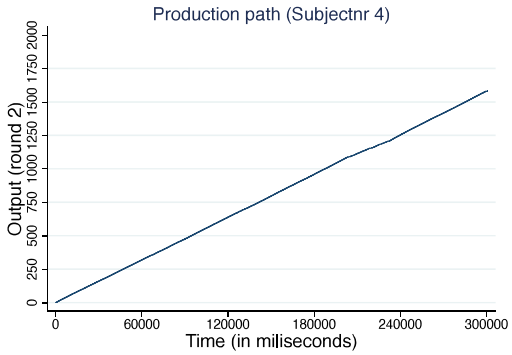
Table A.1: Reference population and the three potential reference workers (in **bold**)

rank	subjectnr	effort1	effort2
1	36	1553	1128
2	25	1488	1474
3	6	1446	1458
4	**4**	**1428**	**1580**
5	13	1415	1048
6	39	1409	1426
7	15	1366	544
8	19	1338	519
9	27	1325	1231
10	16	1307	1338
11	42	1301	1016
12	18	1299	1300
13	55	1284	1244
14	23	1259	861
15	20	1249	1226
16	3	1238	1081
17	51	1231	1326
18	54	1198	1310
19	47	1189	1109
20	8	1189	1133
21	38	1149	1258
22	21	1119	1297
23	56	1111	1402
24	48	1105	257
25	43	1077	1032
26	**46**	**1073**	**1195**
27	45	1062	1254
28	22	984	1139
29	14	968	1095
30	9	951	1126
31	30	950	891
32	2	929	1339
33	26	917	982
34	31	914	1058
35	7	897	1012
36	12	893	861
37	11	851	795
38	53	826	822
39	60	820	1069
40	37	809	1261
41	1	805	825
42	33	798	875
43	35	797	1246
44	59	778	888
45	57	739	853
46	50	707	900
47	34	694	714
48	58	589	528
49	**29**	**584**	**678**
50	41	337	171
51	28	336	333
52	24	250	179
53	52	229	302
54	17	205	174
55	10	139	111
56	49	118	126
57	32	101	0
58	44	2	995
59	40	0	812
60	5	0	944

Figure A.6: Production paths for the high productivity reference worker (HI, subjectnr=4, rank=4)

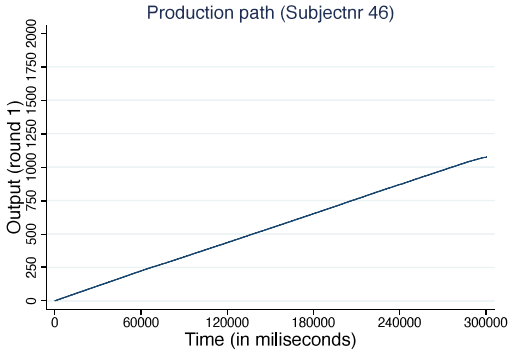


(a) Round 1

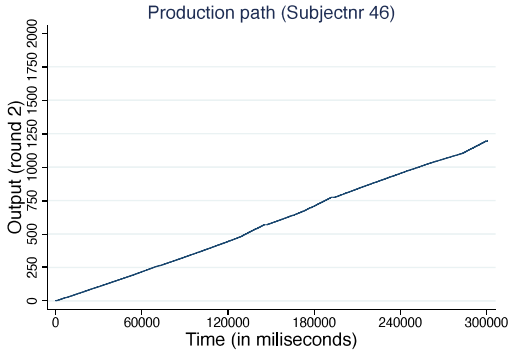


(b) Round 2

Figure A.7: Production paths for the average productivity reference worker (MI, subjectnr=46, rank=26)

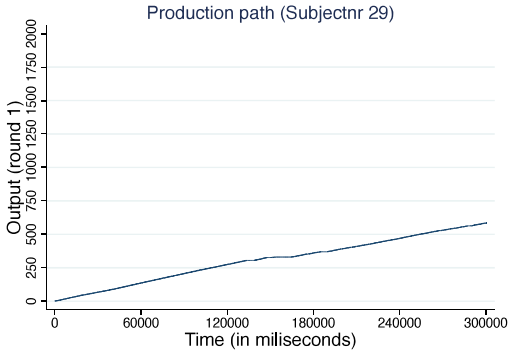


(a) Round 1

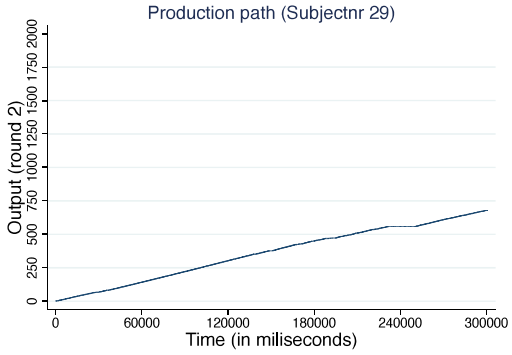


(b) Round 2

Figure A.8: Production paths for the low productivity worker (LO, subjectnr=29, rank=49)



(a) Round 1



(b) Round 2

A.3 Description of the different treatments and sample size

In the Table A.2 below, we depict the key characteristics of each treatment we conducted (whether subjects can compare to a reference worker or not, the matching procedure, and whether subjects were paid a piece-rate on top of their base payment) and the associated sample size. Note that, in the EXRA condition, subjects were randomly (and uniformly) assigned to one of the four sub-conditions: EXRA-NO, EXRA-LO, EXRA-MI, EXRA-HI. We therefore have approximately 500 participants in each of these subconditions.

Table A.2: The key features of the different treatments

Treatment	Comparisons possible	Matching procedure	Piece-rate	Sample size
RANK	No	-	No	1016
EXRA	Yes	Exogenous (Random)	No	2028
ENDO	Yes	Endogenous (Choice)	No	1001
EXBE	Yes	Exogenous (Most motivating)	No	503
RANK$	No	-	Yes	499
ENDO$	Yes	Endogenous (Choice)	Yes	993
EXBE$	Yes	Exogenous (Most motivating)	Yes	492

Notes: "Comparison possible" indicates whether comparisons to a reference worker is possible (Yes) or not (No). "Matching procedure" indicate the process through which workers are matched with a reference worker (if any): 'Random' indicates that workers are randomly assigned to reference workers, 'Choice' indicates that workers can choose a reference worker, 'Non-random and no choice' indicates that workers are forced to compare to reference worker based on a non-random procedure. "Piece-rate" indicates whether workers where paid an additional piece rate on top of their flat payment.

A.4 Descriptive statistics on the population sample

We depict the main descriptive statistics for our study in the Table A.3 below.

Table A.3: Descriptive statistics

	Mean	S.D.	Min	Max	N
Male (=1)	0.4	0.5	0	1	6532
Age	36.2	12.3	8	118	6532
Effort round 1	1042.2	309.3	0	1905	6532
Effort round 2	1173.1	368.5	0	2000	6532
Total Effort (Effort1 + Effort2)	2215.3	620.0	1	3905	6532
Beliefs about own effort (round 1)	617.6	656.3	0	3000	6532
Beliefs about own effort (round 2)	1020.4	424.6	0	3000	6532
Observations	6532				

Notes: Male is a dummy variable which equals one if the subject's gender is male. Age is a continuous variable. Effort in round 1 (2) represents workers' output in round 1 (2). Total Effort is workers' total output across production rounds. Beliefs (about own effort in round 1, and 2) correspond to workers' expectations regarding their own output (winsorized at 3000).

A.5 Balance checks and attrition

In Table A.4, we regress workers' main observable characteristics (effort in round 1, age, male) on a set of dummy variables indicating treatment assignment and a dummy controlling for the timing of the data collection (Wave). The omitted category are participants in the RANK condition. For all three variables, an omnibus test of condition assignment does not reject the null hypothesis of equal observables across conditions (See "Joint F-Test (p-value) at the bottom of the Table). We therefore conclude that our subjects are well randomized into treatments.

Table A.4: Balance test

	effort 1	age	male
	(1)	(2)	(3)
EXRA-HI	17.697	1.186	-0.010
	(19.669)	(0.780)	(0.032)
EXRA-MI	-5.883	1.154	-0.017
	(19.761)	(0.763)	(0.031)
EXRA-LO	24.071	0.824	0.030
	(19.362)	(0.744)	(0.032)
EXRA-NO	15.890	0.966	-0.009
	(19.696)	(0.739)	(0.032)
ENDO	5.998	-0.121	-0.013
	(15.844)	(0.725)	(0.027)
EXBE	-26.867	-0.367	-0.007
	(18.399)	(0.818)	(0.031)
EXBE×$	-10.577	-1.196	0.027
	(19.155)	(0.816)	(0.031)
RANK×$	-20.315	-0.362	-0.010
	(19.363)	(0.815)	(0.031)
ENDO×$	-9.867	-0.225	0.005
	(15.987)	(0.720)	(0.027)
Wave	26.215	0.990	-0.058*
	(18.639)	(0.800)	(0.031)
R^2	0.001	0.001	0.004
Joint F-test (p-value)	0.471	0.679	0.813
Observations	6532	6532	6532

Notes: OLS estimations. The dependent variable is indicated at the top of each column. All the variables are dummies indicating treatment assignment. The omitted category are participants in the RANK condition. "Wave" is a control for whether the data collection took place in wave 1 (EXRA treatments) or in wave 2 (all other treatments). Note that RANK data was collected in both waves. The Joint F-Test is an omnibus test of significance of all the treatment dummies, controlling for the wave. Levels of significance: $*p < 0.1, **p < 0.05, ***p < 0.01$.

In total, 7385 subjects clicked on our HIT. In Table A.5, we regress a dummy variable which equals one if the subject who initially enrolled for the study dropped out of the study before the end on a set of dummies indicating treatment assignment. The regression clearly indicates that attrition is independent of treatment assignment (Joint F-Test: $p = 0.797$).

Table A.5: Attrition

	Attrition (dropped=1)
	(1)
EXRA-HI	-0.003
	(0.019)
EXRA-MI	-0.026
	(0.018)
EXRA-LO	-0.015
	(0.019)
EXRA-NO	-0.006
	(0.019)
ENDO	0.016
	(0.015)
EXBE	0.002
	(0.017)
EXBE×$	-0.001
	(0.017)
RANK×$	0.014
	(0.018)
ENDO×$	0.001
	(0.015)
Wave	-0.031*
	(0.018)
R^2	0.001
Joint F-test (p-value)	0.797
Observations	7385

Notes: OLS estimation. The dependent variable is a dummy which equals one if a subject who initially enrolled for the study (i.e. clicked on the HIT) dropped before the end of the assignment. The different variables indicate the different treatment conditions. The omitted category are participants in the RANK condition. "Wave" is a control for whether the data collection took place in wave 1 (EXRA treatments) or in wave 2 (all other treatments). Note that RANK data was collected in both waves. The Joint F-Test is an omnibus test of significance of all the treatment dummies, controlling for the wave. Levels of significance: $*p < 0.1, **p < 0.05, ***p < 0.01$.

B Additional material related to the estimation of treatment effects

This Appendix contains the material related to the estimation of the treatment effects reported throughout the paper. We start by outlining the empirical strategy. We then present the results of the different estimations, following the structure of the main paper.

B.1 Estimation strategy

We leverage the panel-structure of the data (we observe effort at the individual level in two consecutive periods, i.e. effort1 and effort2). In the simplest case, e.g. when comparing the effect of ranking information (RANK) to the effects of exogenously assigned reference workers (EXRA), we estimate the following model

$$\begin{aligned}\text{effort}_{it} =& \beta_1 \text{Treatment1}_i + \beta_2 \text{Treatment2}_i \\ &+ \beta_3(\text{Treatment1}_i \times \text{P2}_t) + \beta_4(\text{Treatment2}_i \times \text{P2}_t) + \epsilon_{it}\end{aligned}$$

where effort_{it} it the effort of individual i in period t, Treatment1_i and Treatment2_i are individual-specific treatment dummies which take the value of one if the individual is in the respective treatment, and P2_t is a dummy which take the value of one if the observation comes from period 2. The residulas ϵ_{it} are clustered at the individual level.[42]

Our main interest is to compare β_3 and β_4. These two coefficients tell us by how much output *changes* between period 1 and period 2 in the two respective treatments, i.e. it revelas which treatments yields the largest effect. For simplicity, we *only* report these coefficients in the following tables. These treatments effects (and the associated p-values) are also the ones reported in the main text.

In such a model, β_1 and β_2 reveal the period-1 output in the different treatments. Because our treatments are operationalized at the beginning of the second production round and by virtue of randomization, output in round 1 can *not* be affected by the treatments.[43] We therefore do not report these coefficients in the regression tables (they are indicated by the row "Treatment dummies"). However, the main text always refers to the period-specific production levels when discussing treatment effects.

[42] Note that, in all the tables, we also report the estimates of a model that also includes individual-specific controls for age and gender.

[43] Moreover, we have shown in Appendix A.5 that no such differences exist, i.e. that the treatments are well balanced with respect to workers' observable characteristics, including period 1 output.

B.2 The effects of randomly assigned reference workers

B.2.1 Average treatment effects

Following the procedure described above, we estimate the following model:

$$\begin{aligned}\text{effort}_{it} =& \beta_1 \text{EXRA}_i + \beta_2 \text{RANK}_i \\ &+ \beta_3(\text{EXRA}_i \times \text{P2}_t) + \beta_4(\text{RANK}_i \times \text{P2}_t) + \epsilon_{it}\end{aligned}$$

We report the results in the Table B.1 below. "EXRA x P2" shows the motivational effect of being assigned to the EXRA treatment (β_3), i.e. by how much production increases from period 1 to period 2 in the EXRA treatment. Similarly, "RANK x P2" shows the motivational effect of the RANK treatment (β_4), i.e. by how much production increases from period 1 to period 2 in the RANK treatment. The baseline productivity levels (β_1 and β_2 are identical across treatments by virtue of randomization and are therefore not reported, see "Treatment dummies"). This table unambiguously shows that the increase in performance generated by EXRA (+ 117.391 units of output, $p < 0.01$) is almost twice the size of that generated by RANK (+67.021, $p < 0.01$). The changes in output generated by these two treatments are significantly different from each other, as reported by the test at the bottom of the table (Ho: EXRA x P2 = RANK x P2, $p = 0.000$).

All the other Tables are constructed in a similar way. We therefore only display the regression outputs for the remaining estimations.

Table B.1: The effects of randomly assigned reference workers

	Effort	
	(1)	(2)
EXRA x P2	117.391***	117.391***
	(6.593)	(6.594)
RANK x P2	67.021***	67.021***
	(9.466)	(9.467)
Male		40.605***
		(11.302)
Age		-4.339***
		(0.450)
Treatment dummies	Yes	Yes
Ho : EXRA x P2 = RANK x P2	0.000	0.000
R^2	0.908	0.911
Observations	3044	3044

Note: OLS estimations. "EXRA x P2" is a dummy which equals 1 if the participant was assigned to the EXRA treatment. "RANK x P2" is a dummy which equals 1 if the participant was in the RANK treatment. These coefficients indicate by how much effort changed from period 1 to period 2 in the respective treatments. "Treatment dummies" comprises a set of dummy variables capturing treatment-specific period 1 effort. Male and Age are further individual-level controls. P-values of test of equality in coefficients are reported at the bottom of the table. For example, the line "Ho: EXRA x P2 = RANK x P2" provides the p-value of a test of equality between the "EXRA x P2" and the "RANK x P2" coefficients. Levels of significance: $*p < 0.1, **p < 0.05, ***p < 0.01$.

B.2.2 What are the effects of the different randomly assigned reference workers?

Table B.2: The effects of the different randomly assigned reference workers

	Effort		Stress	
	(1)	(2)	(3)	(4)
EXRA-HI x P2	148.040***	148.040***	0.705***	0.705***
	(12.554)	(12.556)	(0.048)	(0.048)
EXRA-MI x P2	125.213***	125.213***	0.577***	0.577***
	(13.515)	(13.517)	(0.047)	(0.047)
EXRA-LO x P2	111.895***	111.895***	0.327***	0.327***
	(11.549)	(11.551)	(0.046)	(0.046)
EXRA-NO x P2	84.086***	84.086***	0.357***	0.357***
	(14.752)	(14.754)	(0.045)	(0.045)
RANK x P2	67.021***	67.021***	0.472***	0.472***
	(9.470)	(9.472)	(0.031)	(0.031)
Male		40.370***		-0.030
		(11.310)		(0.045)
Age		-4.340***		-0.004**
		(0.451)		(0.002)
Treatment dummies	Yes	Yes	Yes	Yes
Ho: EXRA-HI x P2 = EXRA-MI x P2	0.216	0.216	0.056	0.056
Ho: EXRA-HI x P2 = EXRA-LO x P2	0.034	0.034	0.000	0.000
Ho: EXRA-HI x P2 = EXRA-NO x P2	0.001	0.001	0.000	0.000
Ho: EXRA-MI x P2 = EXRA-LO x P2	0.454	0.454	0.000	0.000
Ho: EXRA-MI x P2 = EXRA-NO x P2	0.040	0.040	0.001	0.001
Ho: EXRA-LO x P2 = EXRA-NO x P2	0.138	0.138	0.651	0.651
R^2	0.909	0.911	0.805	0.806
Observations	3044	3044	3044	3044

Note: OLS estimations. "EXRA-HI x P2" is a dummy which equals 1 if the participant was assigned to the EXRA-HI treatment. "EXRA-MI x P2" is a dummy which equals 1 if the participant was in the EXRA-MI treatment. These coefficients indicate by how much effort (resp. stress) changed from period 1 to period 2 in the respective treatments. The remaining interactions variables are defined in a similar way. "Treatment dummies" comprises a set of dummy variables capturing treatment-specific period 1 effort (resp. stress). Male and Age are further individual-level controls. P-values of test of equality in coefficients are reported at the bottom of the table. For example, the line "Ho: EXRA-HI x P2 = EXRA-MI x P2" provides the p-value of a test of equality between the "EXRA-HI x P2" and the "EXRA-MI x P2" coefficients. Levels of significance: $*p < 0.1, **p < 0.05, ***p < 0.01$.

B.2.3 Do the effects of randomly assigned reference workers depend on the characteristics of the observer?

The role played by the performance in round 1 of the observer

We start the heterogeneity analysis at the descriptive level. Figure B.1 depicts the causal effects of the different exogenously assigned reference workers, separately for workers with different performance levels in round 1. Following what we pre-registered, we divide the our sample into the following four subsamples:

a) workers with an output in period 1 that is *lower* than the output in period 1 of the least productive reference worker ($e_1 \leq e_1(LO)$)

b) workers with an output in period 1 that is *higher* than the output in period 1 of the least productive reference worker, but *lower* than the output in period 1 of the average productivity worker ($e_1(LO) \leq e_1 \leq e_1(MI)$)

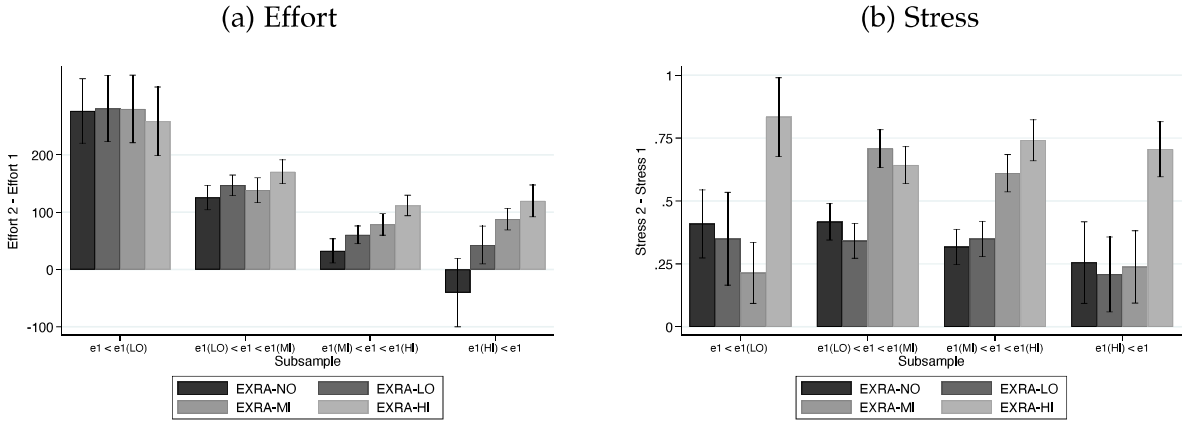
c) workers with an output in period 1 that is *higher* than the output in period 1 of the average productivity reference worker, but *lower* than the output in period 1 of the high productivity worker ($e_1(MI) \leq e_1 \leq e_1(HI)$)

d) workers withan output in period 1 that is *higher* than the output in period 1 of the high productivity worker ($e_1(HI) \leq e_1$)

In line with the overall patterns reported in the main text, Figure B.1a shows that, for most workers, productivity gains between periods 1 and 2 tend to increase in the performance of the reference worker assigned to them. The only exception are workers from subsample a), whose performance in round 1 was lower than the performance in round 1 of the low productivity worker. For them, no clear pattern emerges. For all the workers, while being assigned a more productive reference worker generates a larger increase in productivity, it also generates a larger increase in stress, as documented in Figure B.1b.

These results are corroborated by regression analysis (see Table B.3 and B.4). For all the workers *except* those in the least productive segment of the distribution, the largest increase in performance is achieved by workers who are exogenously assigned to the *most* productive reference worker (EXRA-HI). For example, workers in the third subsample (whose output in round 1 is higher than the output in round 1 of the average reference worker, but lower than the round 1 output of the highly productive reference worker, see column 5 and 6), increase their production by 75.34 units when exogenously assigned to HI ($p < 0.01$), by 41.06 when assigned to MI ($p < 0.1$) and by 29.14 if assigned to LO. While the differences between coefficients are not always significant; the point estimates are always the largest for EXRA-HI in columns 3-8, and the largest for EXRA-NO in columns 1-2. Turning to stress, the regression results are generally consistent with the descriptive evidence: being assigned to the most productive reference worker tends to generate the largest increase in stress (see Table B.4).

Figure B.1: Effects of different exogenously assigned reference worker, by subsample



Note: Panel a) depicts the average change in effort between rounds 1 and 2. Panel b) depicts the average change in stress between rounds 1 and 2. Sample divided into 4 subsamples: a) workers with a production in round 1 that is lower than the production in round 1 of the least productive reference worker ($e_1 \leq e_1(LO)$), b) workers with a production in round 1 that is between the production in round 1 of the least productive reference worker and the average reference worker ($e_1(LO) \leq e_1 \leq e_1(MI)$), c) workers with a production in round 1 that is between the production in round 1 of the average reference worker and the most productive reference worker ($e_1(MI) \leq e_1 \leq e_1(HI)$), d) workers with a production in round 1 that is higher than the most productive reference worker ($e_1(HI) \leq e_1$). Stress levels were measured after each round using the question "On a scale from 1 to 5, how stressed have you been while completing the task?" Answer categories ranged from "Not at all stressed" (1) to "Very stressed" (5).

Table B.3: The effects of exogenously assigned reference workers on effort (by period 1 output)

	e1 < LOW		LOW < e1 < MED		MED < e1 < HI		e1 > HI	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
EXRA-HI x P2	258.676***	258.676***	170.733***	170.733***	111.839***	111.839***	119.843***	119.843***
	(59.776)	(59.896)	(20.743)	(20.752)	(17.809)	(17.816)	(27.926)	(27.983)
EXRA-MI x P2	280.321***	280.321***	138.541***	138.541***	78.704***	78.704***	87.952***	87.952***
	(59.057)	(59.176)	(21.509)	(21.517)	(18.804)	(18.811)	(18.684)	(18.722)
EXRA-LO x P2	281.225***	281.225***	147.194***	147.194***	60.536***	60.536***	42.917	42.917
	(57.609)	(57.725)	(17.715)	(17.722)	(15.605)	(15.611)	(33.025)	(33.093)
EXRA-NO x P2	276.864***	276.864***	125.663***	125.663***	32.642	32.642	-40.383	-40.383
	(56.338)	(56.451)	(21.223)	(21.232)	(21.146)	(21.155)	(59.659)	(59.781)
RANK x P2	216.103***	216.103***	83.500***	83.500***	34.129***	34.129***	5.429	5.429
	(44.919)	(45.009)	(14.475)	(14.481)	(13.107)	(13.112)	(37.578)	(37.654)
Male		-41.427		-17.658		-0.001		8.310
		(27.437)		(10.932)		(9.144)		(20.039)
Age		0.580		-1.517***		-1.032***		0.338
		(0.874)		(0.388)		(0.388)		(0.988)
Treatment dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Ho: EXRA-HI x P2 = EXRA-MI x P2	0.797	0.797	0.282	0.282	0.201	0.201	0.343	0.344
Ho: EXRA-HI x P2 = EXRA-LO x P2	0.786	0.787	0.388	0.389	0.030	0.031	0.077	0.077
Ho: EXRA-HI x P2 = EXRA-NO x P2	0.825	0.825	0.129	0.129	0.004	0.004	0.016	0.016
Ho: EXRA-MI x P2 = EXRA-LO x P2	0.991	0.991	0.756	0.756	0.457	0.457	0.236	0.237
Ho: EXRA-MI x P2 = EXRA-NO x P2	0.966	0.966	0.670	0.670	0.104	0.104	0.041	0.042
Ho: EXRA-LO x P2 = EXRA-NO x P2	0.957	0.957	0.436	0.436	0.289	0.289	0.223	0.224
R^2	0.745	0.747	0.945	0.946	0.972	0.972	0.982	0.982
Observations	255	255	1250	1250	1288	1288	251	251

Note: OLS estimations. "EXRA-HI x P2" is a dummy which equals 1 if the participant was assigned to the EXRA-HI treatment. "EXRA-MI x P2" is a dummy which equals 1 if the participant was in the EXRA-MI treatment. These coefficients indicate by how much effort changed from period 1 to period 2 in the respective treatments. The remaining interactions variables are defined in a similar way. "Treatment dummies" comprises a set of dummy variables capturing treatment-specific period 1 effort. Male and Age are further individual-level controls. P-values of test of equality in coefficients are reported at the bottom of the table. For example, the line "Ho: EXRA-HI x P2 = EXRA-MI x P2" provides the p-value of a test of equality between the "EXRA-HI x P2" and the "EXRA-MI x P2" coefficients. Sample divided into 4 subsamples: a) workers with a production in round 1 that is lower than the production in round 1 of the least productive reference worker ($e_1 \leq e_1(LO)$), b) workers with a production in round 1 that is between the production in round 1 of the least productive reference worker and the average reference worker ($e_1(LO) \leq e_1 \leq e_1(MI)$), c) workers with a production in round 1 that is between the production in round 1 of the average reference worker and the most productive reference worker ($e_1(MI) \leq e_1 \leq e_1(HI)$), d) workers with a production in round 1 that is higher than the most productive reference worker ($e_1(HI) \leq e_1$). Levels of significance: $^*p < 0.1, ^{**}p < 0.05, ^{***}p < 0.01$.

Table B.4: The effects of exogenously assigned reference workers on stress (by period 1 output)

	e1 < LOW		LOW < e1 < MED		MED < e1 < HI		e1 > HI	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
EXRA-HI x P2	0.869***	0.871***	0.643***	0.643***	0.741***	0.741***	0.706***	0.706***
	(0.157)	(0.158)	(0.073)	(0.073)	(0.081)	(0.081)	(0.110)	(0.110)
EXRA-MI x P2	0.214*	0.214*	0.708***	0.708***	0.610***	0.610***	0.238*	0.238*
	(0.122)	(0.122)	(0.075)	(0.075)	(0.074)	(0.074)	(0.144)	(0.144)
EXRA-LO x P2	0.350*	0.350*	0.328***	0.328***	0.349***	0.349***	0.208	0.208
	(0.184)	(0.185)	(0.070)	(0.070)	(0.070)	(0.070)	(0.149)	(0.149)
EXRA-NO x P2	0.409***	0.409***	0.417***	0.417***	0.311***	0.311***	0.255	0.255
	(0.135)	(0.135)	(0.072)	(0.072)	(0.070)	(0.070)	(0.162)	(0.163)
RANK x P2	0.295***	0.295***	0.432***	0.432***	0.521***	0.521***	0.619***	0.619***
	(0.094)	(0.094)	(0.049)	(0.049)	(0.048)	(0.048)	(0.129)	(0.130)
Male		0.154		-0.087		-0.083		-0.056
		(0.156)		(0.072)		(0.067)		(0.172)
Age		0.003		0.001		-0.007**		0.001
		(0.005)		(0.002)		(0.003)		(0.006)
Treatment dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Ho: EXRA-HI x P2 = EXRA-MI x P2	0.001	0.001	0.533	0.533	0.234	0.234	0.010	0.010
Ho: EXRA-HI x P2 = EXRA-LO x P2	0.033	0.033	0.002	0.002	0.000	0.000	0.008	0.008
Ho: EXRA-HI x P2 = EXRA-NO x P2	0.027	0.027	0.028	0.028	0.000	0.000	0.022	0.023
Ho: EXRA-MI x P2 = EXRA-LO x P2	0.539	0.540	0.000	0.000	0.011	0.011	0.886	0.886
Ho: EXRA-MI x P2 = EXRA-NO x P2	0.285	0.286	0.005	0.005	0.003	0.003	0.937	0.937
Ho: EXRA-LO x P2 = EXRA-NO x P2	0.796	0.797	0.376	0.375	0.701	0.701	0.831	0.832
R^2	0.776	0.777	0.798	0.799	0.816	0.817	0.836	0.836
Observations	255	255	1250	1250	1288	1288	251	251

Note: OLS estimations. "EXRA-HI x P2" is a dummy which equals 1 if the participant was assigned to the EXRA-HI treatment. "EXRA-MI x P2" is a dummy which equals 1 if the participant was in the EXRA-MI treatment. These coefficients indicate by how much stress changed from period 1 to period 2 in the respective treatments. The remaining interactions variables are defined in a similar way. "Treatment dummies" comprises a set of dummy variables capturing treatment-specific period 1 stress. Male and Age are further individual-level controls. P-values of test of equality in coefficients are reported at the bottom of the table. For example, the line "Ho: EXRA-HI x P2 = EXRA-MI x P2" provides the p-value of a test of equality between the "EXRA-HI x P2" and the "EXRA-MI x P2" coefficients. Sample divided into 4 subsamples: a) workers with a production in round 1 that is lower than the production in round 1 of the least productive reference worker ($e_1 \leq e_1(LO)$), b) workers with a production in round 1 that is between the production in round 1 of the least productive reference worker and the average reference worker ($e_1(LO) \leq e_1 \leq e_1(MI)$), c) workers with a production in round 1 that is between the production in round 1 of the average reference worker and the most productive reference worker ($e_1(MI) \leq e_1 \leq e_1(HI)$), d) workers with a production in round 1 that is higher than the most productive reference worker ($e_1(HI) \leq e_1$). Levels of significance: $^*p < 0.1, ^{**}p < 0.05, ^{***}p < 0.01$.

The role played by the interaction of the gender of the observer and their performance in round 1

In this Appendix, we further explore heterogeneous responses to the different reference workers by breaking down the sample both by performance in round 1 *and* by gender. Overall, the results are largely consistent with the patterns documented above, i.e. gender is not a key determinant for how participants' productivity respond to the different reference workers. Similarly, male and female participants from different subsamples predominantly react to reference workers in a similar way: the high productivity reference worker is generally the most stressful.

Table B.5: The effects of exogenously assigned reference workers on effort (by period 1 output, male sample)

	e1 < LOW		LOW < e1 < MED		MED < e1 < HI		e1 > HI	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
EXRA-HI x P2	184.762**	184.762**	154.357***	154.357***	97.594***	97.594***	98.125**	98.125**
	(72.543)	(72.683)	(36.936)	(36.956)	(23.223)	(23.232)	(42.817)	(42.881)
EXRA-MI x P2	294.226***	294.226***	115.679***	115.679***	61.424*	61.424*	86.750***	86.750***
	(89.770)	(89.943)	(40.940)	(40.963)	(33.405)	(33.418)	(25.715)	(25.753)
EXRA-LO x P2	317.500***	317.500***	110.487***	110.487***	45.290*	45.290*	42.605	42.605
	(91.937)	(92.114)	(36.543)	(36.563)	(24.478)	(24.487)	(39.511)	(39.570)
EXRA-NO x P2	322.682***	322.682***	140.112***	140.112***	-8.980	-8.980	-84.229	-84.229
	(96.275)	(96.461)	(39.743)	(39.764)	(39.431)	(39.446)	(78.059)	(78.175)
RANK x P2	183.390***	183.390***	62.122**	62.122**	7.749	7.749	30.073	30.073
	(67.710)	(67.841)	(31.397)	(31.414)	(21.690)	(21.698)	(33.312)	(33.361)
Age		1.417		-0.301		-1.520**		2.164
		(1.071)		(0.573)		(0.606)		(1.373)
Treatment dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Ho: EXRA-HI x P2 = EXRA-MI x P2	0.345	0.346	0.483	0.484	0.374	0.374	0.820	0.820
Ho: EXRA-HI x P2 = EXRA-LO x P2	0.259	0.260	0.399	0.399	0.122	0.122	0.342	0.343
Ho: EXRA-HI x P2 = EXRA-NO x P2	0.255	0.256	0.793	0.793	0.020	0.020	0.042	0.042
Ho: EXRA-MI x P2 = EXRA-LO x P2	0.857	0.857	0.925	0.925	0.697	0.697	0.350	0.351
Ho: EXRA-MI x P2 = EXRA-NO x P2	0.829	0.830	0.669	0.669	0.174	0.174	0.039	0.039
Ho: EXRA-LO x P2 = EXRA-NO x P2	0.969	0.969	0.583	0.584	0.243	0.243	0.149	0.150
R^2	0.695	0.697	0.927	0.927	0.963	0.963	0.980	0.980
Observations	135	135	463	463	664	664	174	174

Note: OLS estimations. "EXRA-HI x P2" is a dummy which equals 1 if the participant was assigned to the EXRA-HI treatment. "EXRA-MI x P2" is a dummy which equals 1 if the participant was in the EXRA-MI treatment. These coefficients indicate by how much effort changed from period 1 to period 2 in the respective treatments. The remaining interactions variables are defined in a similar way. "Treatment dummies" comprises a set of dummy variables capturing treatment-specific period 1 effort. Male and Age are further individual-level controls. P-values of test of equality in coefficients are reported at the bottom of the table. For example, the line "Ho: EXRA-HI x P2 = EXRA-MI x P2" provides the p-value of a test of equality between the "EXRA-HI x P2" and the "EXRA-MI x P2" coefficients. Sample divided into 4 subsamples: a) workers with a production in round 1 that is lower than the production in round 1 of the least productive reference worker ($e_1 \leq e_1(LO)$), b) workers with a production in round 1 that is between the production in round 1 of the least productive reference worker and the average reference worker ($e_1(LO) \leq e_1 \leq e_1(MI)$), c) workers with a production in round 1 that is between the production in round 1 of the average reference worker and the most productive reference worker ($e_1(MI) \leq e_1 \leq e_1(HI)$), d) workers with a production in round 1 that is higher than the most productive reference worker ($e_1(HI) \leq e_1$). Levels of significance: $^*p < 0.1, ^{**}p < 0.05, ^{***}p < 0.01$.

Table B.6: The effects of exogenously assigned reference workers on stress (by period 1 output, male sample)

	e1 < LOW		LOW < e1 < MED		MED < e1 < HI		e1 > HI	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
EXRA-HI x P2	0.619***	0.619***	0.595***	0.595***	0.564***	0.564***	0.719***	0.719***
	(0.188)	(0.188)	(0.126)	(0.126)	(0.113)	(0.113)	(0.144)	(0.144)
EXRA-MI x P2	0.194	0.194	0.464***	0.464***	0.566***	0.566***	0.179	0.179
	(0.150)	(0.151)	(0.145)	(0.145)	(0.121)	(0.121)	(0.192)	(0.193)
EXRA-LO x P2	0.150	0.150	0.030	0.029	0.331***	0.331***	0.342**	0.342**
	(0.131)	(0.131)	(0.110)	(0.110)	(0.095)	(0.095)	(0.162)	(0.162)
EXRA-NO x P2	0.273	0.273	0.337***	0.337***	0.346***	0.346***	0.429**	0.429**
	(0.199)	(0.199)	(0.117)	(0.117)	(0.105)	(0.105)	(0.198)	(0.198)
RANK x P2	0.195	0.195	0.324***	0.324***	0.510***	0.510***	0.585***	0.585***
	(0.142)	(0.142)	(0.082)	(0.082)	(0.069)	(0.069)	(0.144)	(0.145)
Age		0.007		0.004		-0.006		-0.004
		(0.006)		(0.004)		(0.004)		(0.009)
Treatment dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Ho: EXRA-HI x P2 = EXRA-MI x P2	0.079	0.080	0.496	0.496	0.994	0.994	0.026	0.026
Ho: EXRA-HI x P2 = EXRA-LO x P2	0.042	0.043	0.001	0.001	0.113	0.114	0.083	0.084
Ho: EXRA-HI x P2 = EXRA-NO x P2	0.208	0.209	0.133	0.134	0.155	0.156	0.237	0.238
Ho: EXRA-MI x P2 = EXRA-LO x P2	0.827	0.828	0.018	0.018	0.127	0.127	0.516	0.517
Ho: EXRA-MI x P2 = EXRA-NO x P2	0.751	0.752	0.496	0.496	0.169	0.170	0.367	0.367
Ho: EXRA-LO x P2 = EXRA-NO x P2	0.607	0.608	0.056	0.056	0.916	0.914	0.736	0.736
R^2	0.785	0.786	0.793	0.794	0.810	0.811	0.834	0.834
Observations	135	135	463	463	664	664	174	174

Note: OLS estimations. "EXRA-HI x P2" is a dummy which equals 1 if the participant was assigned to the EXRA-HI treatment. "EXRA-MI x P2" is a dummy which equals 1 if the participant was in the EXRA-MI treatment. These coefficients indicate by how much stress changed from period 1 to period 2 in the respective treatments. The remaining interactions variables are defined in a similar way. "Treatment dummies" comprises a set of dummy variables capturing treatment-specific period 1 stress. Male and Age are further individual-level controls. P-values of test of equality in coefficients are reported at the bottom of the table. For example, the line "Ho: EXRA-HI x P2 = EXRA-MI x P2" provides the p-value of a test of equality between the "EXRA-HI x P2" and the "EXRA-MI x P2" coefficients. Sample divided into 4 subsamples: a) workers with a production in round 1 that is lower than the production in round 1 of the least productive reference worker ($e_1 \leq e_1(LO)$), b) workers with a production in round 1 that is between the production in round 1 of the least productive reference worker and the average reference worker ($e_1(LO) \leq e_1 \leq e_1(MI)$), c) workers with a production in round 1 that is between the production in round 1 of the average reference worker and the most productive reference worker ($e_1(MI) \leq e_1 \leq e_1(HI)$), d) workers with a production in round 1 that is higher than the most productive reference worker ($e_1(HI) \leq e_1$). Levels of significance: $^*p < 0.1, ^{**}p < 0.05, ^{***}p < 0.01$.

Table B.7: The effects of exogenously assigned reference workers on effort (by period 1 output, female sample)

	e1 < LOW		LOW < e1 < MED		MED < e1 < HI		e1 > HI	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
EXRA-HI x P2	355.687***	355.687***	181.651***	181.651***	125.673***	125.673***	156.421***	156.421***
	(96.913)	(97.124)	(24.344)	(24.351)	(26.930)	(26.941)	(19.338)	(19.405)
EXRA-MI x P2	263.080***	263.080***	153.904***	153.904***	93.711***	93.711***	90.357***	90.357***
	(73.917)	(74.079)	(23.212)	(23.219)	(19.843)	(19.851)	(23.452)	(23.534)
EXRA-LO x P2	244.950***	244.950***	168.654***	168.654***	82.776***	82.776***	44.100	44.100
	(70.542)	(70.695)	(18.061)	(18.067)	(13.905)	(13.911)	(54.378)	(54.568)
EXRA-NO x P2	231.045***	231.045***	115.950***	115.950***	70.514***	70.514***	87.500**	87.500**
	(59.151)	(59.280)	(23.464)	(23.472)	(17.981)	(17.989)	(38.473)	(38.608)
RANK x P2	252.351***	252.351***	93.854***	93.854***	64.152***	64.152***	-40.500	-40.500
	(58.851)	(58.979)	(15.219)	(15.224)	(13.082)	(13.087)	(89.532)	(89.844)
Age		-0.645		-2.493***		-0.476		-1.937
		(1.336)		(0.471)		(0.457)		(1.294)
Treatment dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Ho: EXRA-HI x P2 =EXRA-MI x P2	0.449	0.450	0.410	0.410	0.340	0.340	0.033	0.033
Ho: EXRA-HI x P2 = EXRA-LO x P2	0.357	0.358	0.668	0.668	0.157	0.158	0.055	0.056
Ho: EXRA-HI x P2 = EXRA-NO x P2	0.275	0.276	0.052	0.052	0.089	0.089	0.114	0.115
Ho: EXRA-MI x P2 = EXRA-LO x P2	0.859	0.860	0.616	0.616	0.652	0.652	0.437	0.439
Ho: EXRA-MI x P2 = EXRA-NO x P2	0.736	0.736	0.251	0.251	0.387	0.387	0.950	0.950
Ho: EXRA-LO x P2 = EXRA-NO x P2	0.880	0.880	0.075	0.076	0.590	0.590	0.517	0.518
R^2	0.809	0.810	0.957	0.958	0.981	0.982	0.987	0.988
Observations	120	120	787	787	624	624	77	77

Note: OLS estimations. "EXRA-HI x P2" is a dummy which equals 1 if the participant was assigned to the EXRA-HI treatment. "EXRA-MI x P2" is a dummy which equals 1 if the participant was in the EXRA-MI treatment. These coefficients indicate by how much effort changed from period 1 to period 2 in the respective treatments. The remaining interactions variables are defined in a similar way. "Treatment dummies" comprises a set of dummy variables capturing treatment-specific period 1 effort. Male and Age are further individual-level controls. P-values of test of equality in coefficients are reported at the bottom of the table. For example, the line "Ho: EXRA-HI x P2 = EXRA-MI x P2" provides the p-value of a test of equality between the "EXRA-HI x P2" and the "EXRA-MI x P2" coefficients. Sample divided into 4 subsamples: a) workers with a production in round 1 that is lower than the production in round 1 of the least productive reference worker ($e_1 \leq e_1(LO)$), b) workers with a production in round 1 that is between the production in round 1 of the least productive reference worker and the average reference worker ($e_1(LO) \leq e_1 \leq e_1(MI)$), c) workers with a production in round 1 that is between the production in round 1 of the average reference worker and the most productive reference worker ($e_1(MI) \leq e_1 \leq e_1(HI)$), d) workers with a production in round 1 that is higher than the most productive reference worker ($e_1(HI) \leq e_1$). Levels of significance: $^*p < 0.1, ^{**}p < 0.05, ^{***}p < 0.01$.

Table B.8: The effects of exogenously assigned reference workers on stress (by period 1 output, female sample)

	e1 < LOW		LOW < e1 < MED		MED < e1 < HI		e1 > HI	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
EXRA-HI x P2	1.208***	1.209***	0.675***	0.675***	0.913***	0.913***	0.684***	0.684***
	(0.250)	(0.251)	(0.089)	(0.089)	(0.115)	(0.115)	(0.174)	(0.174)
EXRA-MI x P2	0.240	0.240	0.872***	0.872***	0.649***	0.649***	0.357*	0.357*
	(0.203)	(0.203)	(0.076)	(0.076)	(0.091)	(0.091)	(0.199)	(0.200)
EXRA-LO x P2	0.550	0.550	0.500***	0.500***	0.376***	0.376***	-0.300	-0.300
	(0.343)	(0.344)	(0.088)	(0.088)	(0.103)	(0.103)	(0.330)	(0.331)
EXRA-NO x P2	0.545***	0.545***	0.471***	0.471***	0.279***	0.279***	-0.250	-0.250
	(0.183)	(0.183)	(0.091)	(0.091)	(0.095)	(0.095)	(0.216)	(0.217)
RANK x P2	0.405***	0.405***	0.484***	0.484***	0.533***	0.533***	0.682**	0.682**
	(0.120)	(0.121)	(0.060)	(0.060)	(0.067)	(0.067)	(0.262)	(0.263)
Age		-0.003		-0.002		-0.007*		0.004
		(0.008)		(0.003)		(0.004)		(0.008)
Treatment dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Ho: EXRA-HI x P2 =EXRA-MI x P2	0.003	0.003	0.091	0.092	0.072	0.072	0.219	0.221
Ho: EXRA-HI x P2 = EXRA-LO x P2	0.124	0.125	0.163	0.163	0.001	0.001	0.010	0.010
Ho: EXRA-HI x P2 = EXRA-NO x P2	0.035	0.035	0.109	0.110	0.000	0.000	0.001	0.001
Ho: EXRA-MI x P2 = EXRA-LO x P2	0.438	0.439	0.001	0.001	0.047	0.047	0.092	0.093
Ho: EXRA-MI x P2 = EXRA-NO x P2	0.266	0.267	0.001	0.001	0.005	0.005	0.042	0.043
Ho: EXRA-LO x P2 = EXRA-NO x P2	0.991	0.991	0.817	0.817	0.488	0.488	0.899	0.900
R^2	0.770	0.771	0.803	0.803	0.824	0.825	0.852	0.853
Observations	120	120	787	787	624	624	77	77

Note: OLS estimations. "EXRA-HI x P2" is a dummy which equals 1 if the participant was assigned to the EXRA-HI treatment. "EXRA-MI x P2" is a dummy which equals 1 if the participant was in the EXRA-MI treatment. These coefficients indicate by how much stress changed from period 1 to period 2 in the respective treatments. The remaining interactions variables are defined in a similar way. "Treatment dummies" comprises a set of dummy variables capturing treatment-specific period 1 stress. Male and Age are further individual-level controls. P-values of test of equality in coefficients are reported at the bottom of the table. For example, the line "Ho: EXRA-HI x P2 = EXRA-MI x P2" provides the p-value of a test of equality between the "EXRA-HI x P2" and the "EXRA-MI x P2" coefficients. Sample divided into 4 subsamples: a) workers with a production in round 1 that is lower than the production in round 1 of the least productive reference worker ($e_1 \leq e_1(LO)$), b) workers with a production in round 1 that is between the production in round 1 of the least productive reference worker and the average reference worker ($e_1(LO) \leq e_1 \leq e_1(MI)$), c) workers with a production in round 1 that is between the production in round 1 of the average reference worker and the most productive reference worker ($e_1(MI) \leq e_1 \leq e_1(HI)$), d) workers with a production in round 1 that is higher than the most productive reference worker ($e_1(HI) \leq e_1$). Levels of significance: $^*p < 0.1, ^{**}p < 0.05, ^{***}p < 0.01$.

B.3 The effects of non-random assignment mechanisms

Table B.9: The effects of endogenously chosen reference workers and of targeted exogenous matching

	Effort		Stress	
	(1)	(2)	(3)	(4)
RANK x P2	67.021***	67.021***	0.472***	0.472***
	(9.466)	(9.467)	(0.031)	(0.031)
EXRA x P2	117.391***	117.391***	0.492***	0.492***
	(6.593)	(6.594)	(0.023)	(0.023)
ENDO x P2	137.795***	137.795***	0.647***	0.647***
	(7.716)	(7.717)	(0.031)	(0.031)
EXBE x P2	146.012***	146.012***	0.783***	0.783***
	(12.142)	(12.143)	(0.052)	(0.052)
Male		45.426***		-0.056
		(9.255)		(0.037)
Age		-4.704***		-0.004***
		(0.380)		(0.001)
Treatment dummies	Yes	Yes	Yes	Yes
Ho: EXRA x P2 = RANK x P2	0.000	0.000	0.612	0.612
Ho: EXRA x P2 = ENDO x P2	0.044	0.044	0.000	0.000
Ho: EXRA x P2 = EXBE x P2	0.038	0.038	0.000	0.000
Ho: ENDO x P2 = EXBE x P2	0.568	0.568	0.024	0.024
R^2	0.912	0.915	0.806	0.806
Observations	4548	4548	4548	4548

Note: OLS estimations. "RANK x P2" is a dummy which equals 1 if the participant was assigned to the RANK treatment. "EXRA x P2" is a dummy which equals 1 if the participant was in the EXRA treatment. These coefficients indicate by how much effort (resp. stress) changed from period 1 to period 2 in the respective treatments. The remaining interactions variables are defined in a similar way. "Treatment dummies" comprises a set of dummy variables capturing treatment-specific period 1 effort (resp. stress). Male and Age are further individual-level controls. P-values of test of equality in coefficients are reported at the bottom of the table. For example, the line "Ho: EXRA x P2 = RANK x P2" provides the p-value of a test of equality between the "EXRA x P2" and the "RANK x P2" coefficients. Levels of significance: $*p < 0.1, **p < 0.05, ***p < 0.01$.

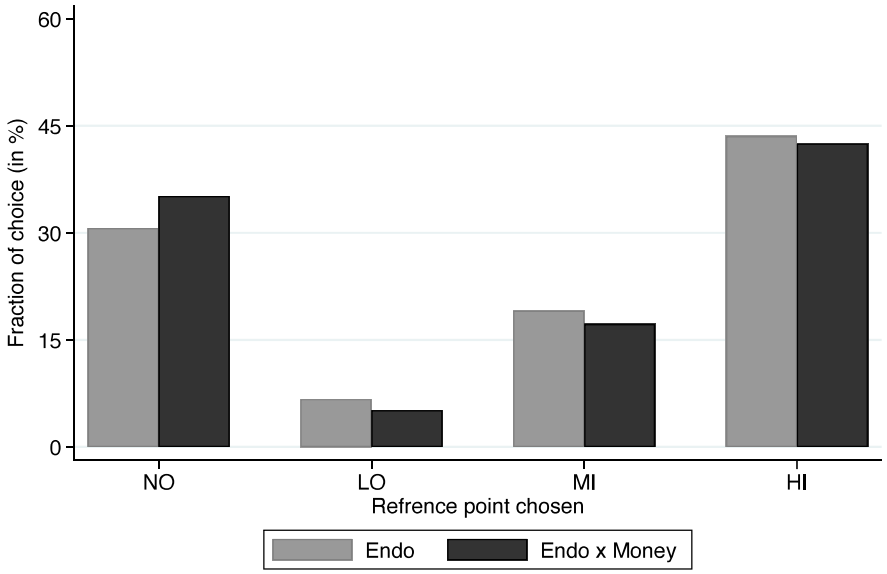
B.4 Benchmarking and robustness

Table B.10: The effects of monetary incentives and social comparisons

	Effort				Stress			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Piece-rate (pooled) x P2	170.147***	170.147***			0.685***	0.685***		
	(5.912)	(5.913)			(0.023)	(0.023)		
Flat wage (pooled) x P2	110.901***	110.901***			0.604***	0.604***		
	(5.508)	(5.509)			(0.021)	(0.021)		
RANK x P2			67.021***	67.021***			0.472***	0.472***
			(9.469)	(9.470)			(0.031)	(0.031)
Rank$ x P2			144.778***	144.778***			0.577***	0.577***
			(11.773)	(11.774)			(0.041)	(0.041)
ENDO x P2			137.795***	137.795***			0.647***	0.647***
			(7.718)	(7.718)			(0.031)	(0.031)
Endo$ x P2			170.633***	170.633***			0.677***	0.677***
			(8.469)	(8.469)			(0.033)	(0.033)
EXBE x P2			146.012***	146.012***			0.783***	0.783***
			(12.145)	(12.146)			(0.052)	(0.052)
EXBE$ x P2			194.896***	194.896***			0.811***	0.811***
			(11.473)	(11.474)			(0.048)	(0.048)
Male		64.297***		65.168***		-0.078**		-0.077**
		(9.252)		(9.232)		(0.037)		(0.037)
Age		-5.357***		-5.365***		-0.006***		-0.006***
		(0.389)		(0.391)		(0.002)		(0.002)
Treatment dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Ho: Flat wage x P2 = Piece-rate x P2	0.000	0.000			0.009	0.009		
Ho: RANK$ x P2 = RANK x P2			0.000	0.000			0.043	0.043
Ho: RANK$ x P2 = ENDO x P2			0.620	0.620			0.173	0.173
Ho: RANK$ x P2 = EXBE x P2			0.942	0.942			0.002	0.002
Ho: RANK$ x P2 = EXBE$ x P2			0.002	0.002			0.000	0.000
Ho: RANK$ x P2 = ENDO$ x P2			0.075	0.075			0.058	0.058
Ho: EXBE$ x P2 = ENDO$ x P2			0.089	0.089			0.023	0.023
R2	0.917	0.921	0.917	0.921	0.804	0.805	0.805	0.805
Observations	4504	4504	4504	4504	4504	4504	4504	4504

Note: OLS estimations. "Piece-rate (pooled) x P2" is a dummy which equals 1 if the participant was in one of the treatment that offered a piece-rate in round 2. "Flat wage (pooled) x P2" is a dummy which equals 1 if the participant was in one of the treatment that did *not* offer a piece-rate in round 2. "RANK x P2" is a dummy which equals 1 if the participant was assigned to the RANK treatment. These coefficients indicate by how much effort (resp. stress) changed from period 1 to period 2 in the respective treatments. The remaining interactions variables are defined in a similar way. "Treatment dummies" comprises a set of dummy variables capturing treatment-specific period 1 effort (resp. stress). Male and Age are further individual-level controls. P-values of test of equality in coefficients are reported at the bottom of the table. For example, the line "Ho: RANK$ x P2 = RANK x P2" provides the p-value of a test of equality between the "RANK$ x P2" and the "RANK x P2" coefficients. Levels of significance: $*p < 0.1, **p < 0.05, ***p < 0.01$.

Figure B.2: The effects of monetary incentives on the distribution of chosen reference workers



Note: Grey bars represent the distribution of choices for the different reference workers in ENDO. Black bars represent the distribution of choices in ENDO×$.

B.5 Exogenously assigning workers to their predicted most motivating reference worker (EXBE)

In the EXBE treatment, workers are exogenously assigned to the reference worker that is predicted to be the most motivating for them, conditional on their observable characteristics (output in round 1 and gender). We use the point estimates discussed in Appendix B.2.3 (Tables B.5 and B.7) as a basis for our predictions.

Our rule for this tailored exogenous matching is therefore:

- If the participant has an input in period 1 that *exceeds* the period 1 output of the **low** productivity reference worker (91.65% of the workers in EXBE), then this participant is assigned to the high productivity reference worker (HI).
- If the participant has an input in period 1 that is *lower* than the period 1 output of the **low** productivity reference worker (8.35% of the workers in EXBE), then this participant is assigned to no reference worker (NO).

All participants in EXBE are assigned to their reference worker according to this rule. Note that the rule applies both to male and female workers as the heterogeneity analysis discussed in Appendix B.2.3 did *not* reveal any gender differences in participants' responses to the different exogenously assigned reference workers.

B.6 The effects of social vs. non-social comparisons (PACE)

In this Appendix, we describe our second pre-registered study aimed at comparing the effects of social and non-social reference points.[44] In this experiment, 500 participants are randomly assigned to the EXRA-HI treatment, while another 500 participants are randomly assigned to a *non-social* "pacemaker" condition (PACE-HI). The EXRA-HI condition is exactly identical to the one implemented in the main study. Subjects in the pacemaker condition are informed that they might see a pacemaker whose speed is randomly determined.[45] The PACE-HI condition differs from the EXRA-HI condition in that participants are not provided with any information about the performance of peers, but are instead presented with a non-social pacemaker whose speed is set such that it reaches exactly the same number of points as the reference worker in the EXRA-HI treatment. Just like in our social treatments, the non-social goal in PACE-HI is operationalized as a growing vertical bar.

These two treatments allow us to compare the effects of social and non-social goals. If social and non-social comparisons are equally motivating, one would expect PACE-HI and EXRA-HI to generate a same-sized increase in performance. However, if social comparisons are more motivating than arbitrarily set goals, one would expect the performance increase to be larger in EXRA-HI.

In Table B.11, we depict the main descriptives. Note that effort in round 1 in this follow-up experiment (mean=1047.7) is remarkably similar to the effort in round 1 in the main study (mean=1042.2, see Table A.3), indicating that no fundamental change in subjects' ability to complete the task occurred across the two studies. Columns 1-2 of Table B.12 show that the motivational spillovers are much larger in the EXRA-HI condition (+123 points, $p < 0.01$) than in the PACE-HI condition (+55 points, $p < 0.01$)—with the two coefficients being highly significantly different from each other ($p < 0.01$). Turning to workers' perceptions, participants in the EXRA-HI condition report being substantially more stressed ($p < 0.01$, columns 3-4 of Table B.12) and more nervous ($p < 0.01$, column 1 of Table B.13) than those in PACE-HI. In addition, participants in EXRA-HI are also more likely to report that the comparison i) motivated them (column 2 of Table B.13, $p < 0.01$), ii) generated a greater feeling of competition (column 3 of Table B.13, $p < 0.01$), and iii) positively affected their performance (column 4 of Table B.13, $p < 0.01$).

[44] This study was pre-registred as trial 137539 on AsPredicted.org and was conducted on Prolific in July 2023.

[45] In order not to deceive subjects, we also assign some subjects to a slow pacemaker condition and a condition without pacemaker. These observations are irrelevant for our analysis and we therefore don't discuss them here (as we pre-registered).

Table B.11: Descriptive statistics

	Mean	S.D.	Min	Max	N
Male	0.6	0.5	0	1	1000
Age	41.4	13.4	18	99	1000
Effort round 1	1047.7	346.0	0	2000	1000
Effort round 2	1148.6	408.7	0	2000	1000
Total effort (Effort1 + Effort2)	2196.3	708.6	2	4000	1000
Beliefs about own effort (round 1)	577.1	613.5	0	3000	1000
Beliefs about own effort (round 2)	1018.8	478.7	0	3000	1000
Observations	1000				

Notes: Male is a dummy variable which equals one if the subject's gender is male. Age is a continuous variable. Effort in round 1 (2) represents workers' output in round 1 (2). Total Effort is workers' total output across production rounds. Beliefs (about own effort in round 1, and 2) correspond to workers' expectations regarding their own output (winsorized at 3000).

Table B.12: The effects of social vs. non-social comparisons

	Effort		Stress	
	(1)	(2)	(3)	(4)
PACE-HI x P2	64.876***	64.876***	0.582***	0.582***
	(11.806)	(11.812)	(0.047)	(0.047)
EXRA-HI x P2	137.046***	137.046***	0.764***	0.764***
	(11.905)	(11.911)	(0.049)	(0.049)
Male		92.097***		-0.100
		(21.718)		(0.079)
Age		-5.155***		-0.005
		(0.730)		(0.003)
Treatment dummies	Yes	Yes	Yes	Yes
Ho : EXRA-HI x P2 = PACE-HI x P2	0.000	0.000	0.008	0.008
R^2	0.895	0.900	0.802	0.803
Observations	1000	1000	1000	1000

Note: OLS estimations."PACE-HI x P2" is a dummy which equals 1 if the participant was in the PACE-HI treatment (non-social pacemaker treatment). "EXRA-HI x P2" is a dummy which equals 1 if the participant was assigned to the EXRA-HI treatment (social comparison). These coefficients indicate by how much effort (resp. stress) changed from period 1 to period 2 in the respective treatments. The remaining interactions variables are defined in a similar way. "Treatment dummies" comprises a set of dummy variables capturing treatment-specific period 1 effort (resp. stress). Male and Age are further individual-level controls. P-values of test of equality in coefficients are reported at the bottom of the table. For example, the line "Ho: EXRA-HI x P2 = PACE-HI x P2" provides the p-value of a test of equality between the "EXRA-HI x P2" and the "PACE-HI x P2" coefficients. Levels of significance: $*p < 0.1, **p < 0.05, ***p < 0.01$.

Table B.13: The effects of social vs. non-social comparisons on workers' perceptions

	(1) Nervous	(2) Motivating	(3) Competition	(4) Performance
Social comparison (EXRA-HI)	0.194**	0.738***	0.231**	0.827***
	(0.083)	(0.168)	(0.091)	(0.161)
Male	-0.121	0.362**	0.200**	0.493***
	(0.084)	(0.169)	(0.092)	(0.163)
Age	-0.019***	-0.020***	-0.028***	-0.014**
	(0.003)	(0.006)	(0.003)	(0.006)
Constant	3.127***	1.761***	4.269***	1.141***
	(0.152)	(0.305)	(0.160)	(0.299)
R^2	0.043	0.036	0.080	0.043
Observations	1000	1000	1000	1000

Note: OLS estimations."Social comprisons (EXRA-HI)" is a dummy which equals 1 if the participant was in the EXRA-HI treatment. Omitted category are participants in the non-social PACE-HI condition. "Nervous" measures whether observing the performance of the reference worker (respectively the pacemaker) made subjects nervous, on a scale from 1 (not at all nervous) to 5 (very nervous). "Motivating" measures how motivating it was for subjects to observe the reference worker (respectively the pacemaker), on a scale from -5 (very discouraging) to +5 (very motivating). "Competition" measures the extent to which subjects felt in competition with the reference worker (respectively the pacemaker), on a scale from 1 (no competition at all) to 5 (very high competition). "Performance" measures subjects' subjective impression that the reference worker (resp. the pacemaker) had on their performance, on a scale from -5 (very negative) to +5 (very positive). Levels of significance: *$p < 0.1$,**$p < 0.05$,***$p < 0.01$.

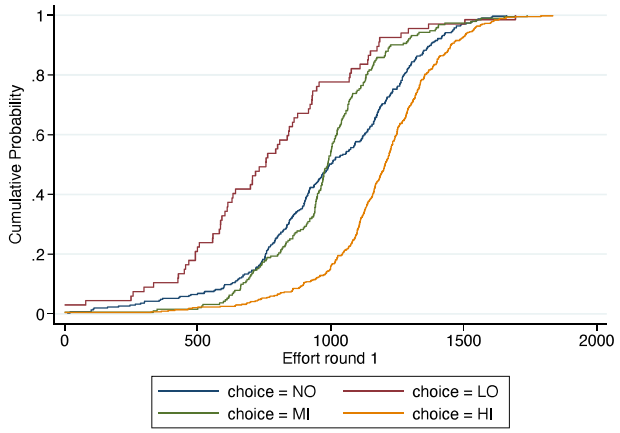
C Additional material related to the analysis of endogenously chosen reference workers

Figure C1a) depicts the cumulative distribution of period 1 effort, conditional on the chosen reference worker. It clearly shows that the distribution of effort in round 1 of workers who choose to compare to the high productivity reference worker (HI) dominates the distributions of workers who choose the average (MI) or the low productivity reference worker (LO). Similarly, the distribution of those who choose to compare to the average productivity reference worker (MI) dominates the distribution of those who choose to compare to the low productivity reference worker (LO). In contrast, the distribution of effort 1 of workers who choose *not* to compare to a reference worker lies in between the distribution of those who compare to the high (HI) and those who compare to the low productivity reference worker (LO). We depict the corresponding density functions in Figure C1b).

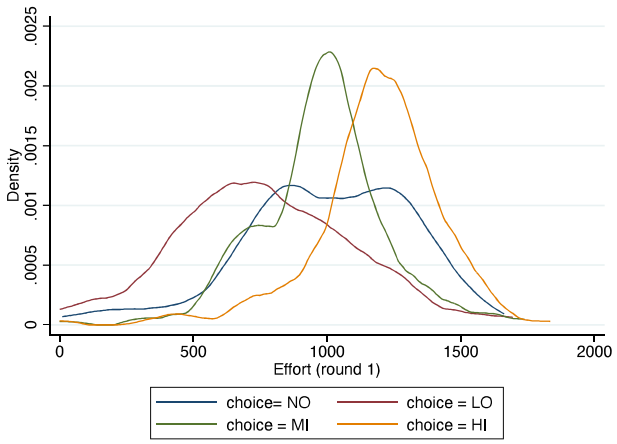
Figure C2 depicts the aggregate distribution of chosen reference workers, separately by gender. While small differences in proportions exist, the overall choice patterns are qualitatively similar: both gender predominantly prefer to compare to the most productive reference worker; the second largest category consists of workers who choose *not* to compare to a reference worker, and the remaining workers compare to either the low or the average productivity reference worker. These qualitative patterns are confirmed by a χ^2 test, which cannot reject the null hypothesis of equal distributions at conventional significance levels ($p = 0.07$).

Figure C1: Distribution of effort in round 1, conditional on chosen reference worker

(a) cumulative distribution

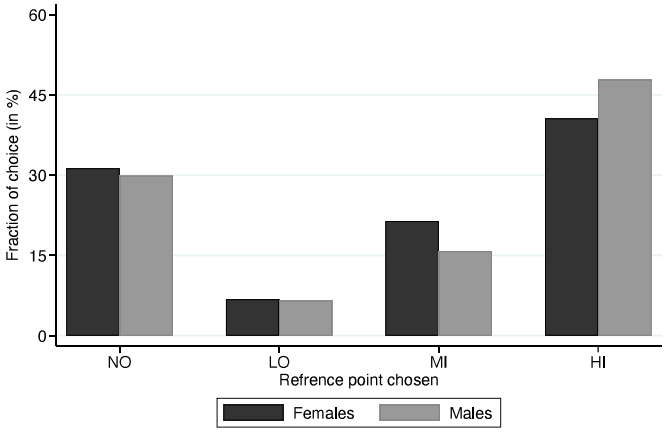


(b) density function



Notes: Distribution of effort in round 1 of workers in the ENDO condition, conditional on chosen reference worker. Panel a) depicts the respective cumulative distributions. Panel b) depicts the corresponding density functions. "NO" corresponds to the distribution of workers who chose *not* to compare to a reference worker. LO (MI, HI) corresponds to the distribution of workers who chose to compare to the low (average, high) productivity reference worker.

Figure C2: Distribution of chosen reference worker (by gender)



Notes: Bars represent the proportion for the four available alternatives, separately by gender. "NO" corresponds to the proportion of workers who chose *not* to compare to a reference worker. LO (MI, HI) corresponds to the proportion of workers who chose to compare to the low (average, high) productivity reference worker.

D Additional material related to the identifications of workers' choice motives

Participants who were given the possibility to choose whom to compare to (ENDO and ENDO×$ conditions) were asked to explain their decision in an open-text format at the end of the study.[46] To identify workers' chief motives and concerns when deciding which reference worker to pick, we hired three independent raters to code participants' answers. Raters were given the following list of nine possible motives (which we identified through focus groups) that could explain workers' choices, along with some examples:

1. Motivation/productivity (e.g. *"To motivate myself"*, *"To push me to go faster"*, *"To help me reach a better score"*)
2. Stress avoidance (e.g. *"I did not want to feel stressed"*, *"It would have been stressful"*, *"It would make me anxious"*)
3. Feel good about self (e.g. *"I compared to this person because he was worse than me"*)
4. Curiosity (e.g. *"I was curious to see how fast/slow he would go"*)
5. Don't care about observing any RP (e.g. *"It didn't matter to see anyone"*)
6. Distraction (e.g. *"I didn't want to get distracted"*)
7. Closest to me (e.g. *"I picked him because he was close to my performance"*)
8. Other (e.g. *"Any answer that cannot be rated using the categories listed above"*)

Each rater was then asked to assign up to three different motives to each answer (i.e. to each worker). The raters were told that they did not need to assign three motives to each answer, i.e. if only one (or two) motive(s) is (are) applicable, they were instructed to leave the remaining motives blank. If an answer could not be categorized, raters were instructed to assign it to the category "Other." For example, a rater could have assigned the answer "*I chose to compare to this reference worker because it was the closest to me and I thought it would motivate me*" both to the category "Motivation" and to the category "Closest to me."

With this procedure, we obtain a maximum of nine different motives (3 raters × 3 possible motives) per rated answer. We aggregate these assessments at the worker-level by extracting the modal motive, i.e. the motive that is most often identified across raters. In order to be able to cleanly interpret these choice motives, the observations for which there is no unique mode are ignored for this analysis.

[46] Participants who decided not to compare to a reference worker were asked the question "*In the previous round, you decided not to observe a reference participant. Please indicate in a few sentences why you made this choice.*" Participants who decided to compare to a reference worker were asked the question "*In the previous round, you observed the performance of the reference participant who ranked XXXth. Please indicate in a few sentences why you have chosen to observe the performance of this participant.*"

E Additional analysis: satisfaction and perceived task difficulty

As we discuss in the paper (see Section 3 on experimental design), we also collected data on satisfaction (*"How satisfied are you with your performance? [1. Not at all satisfied, ..., 5. Very satisfied]"*) as well as percieved task difficulty (*"On a scale from 1 to 5, how difficult did you find the task? [1. Not at all difficult, ..., 5. Very difficult]"*) in addition to the perceived stress that we extensively discuss in the paper. These questions were also asked following each production round. For transparency, we report the effects of the different treatments on these two variables in this Appendix. The Tables are presented in the same order as those for effort and stress (Appendix B). We briefly summarize the main results below. Overall, they are largely consistent with the results on effort and stress documented in the main body of the paper.

Satisfaction Satisfaction about own output generally decreases between rounds in all the treatments, with the largest average decrease in satisfaction reported in the EXBE treatment (see Table E.5 and E.6). In general, being randomly assigned to a very productive reference worker generates a larger decrease in satisfaction than random assignment to an average or low productivity reference worker (see Tables E.2 and E.3). In some cases, being randomly assigned to a *less* productive reference worker increases satisfaction (see e.g. the coefficients "EXRA-MI x P2" in columns 7-8 of Table E.3). Overall, *these results are largely consistent with the stress results presented in the paper*: the treatments that generate the largest *increase* in stress tend to also generate the largest *decrease* in satisfaction.

Perceived task difficulty Being exposed to a reference worker generally increases the perceived task difficulty—consistent with our subjects actively comparing with their reference worker. In general, higher productivity reference worker generate a larger increase in perceived task difficulty than lower productivity reference workers (see Tables E.2 and E.3), with the largest increases in perceived task difficulty being reported in the EXBE condition (see Table E.5 and E.6).

Table E.1: The effects of randomly assigned reference workers

	Satisfaction		Difficulty	
	(1)	(2)	(3)	(4)
EXRA x P2	-0.081***	-0.082***	0.493***	0.493***
	(0.025)	(0.025)	(0.023)	(0.023)
RANK x P2	-0.180***	-0.180***	0.467***	0.467***
	(0.035)	(0.035)	(0.032)	(0.032)
Male		-0.066**		0.009
		(0.033)		(0.046)
Age		0.004***		-0.000
		(0.001)		(0.002)
Treatment dummies	Yes	Yes	Yes	Yes
Ho : EXRA x P2 = RANK x P2	0.021	0.021	0.497	0.497
R^2	0.935	0.936	0.810	0.810
Observations	3044	3044	3044	3044

Note: OLS estimations. "EXRA x P2" is a dummy which equals 1 if the participant was assigned to the EXRA treatment. "RANK x P2" is a dummy which equals 1 if the participant was in the RANK treatment. These coefficients indicate by how much satisfaction (resp. perceived difficulty) changed from period 1 to period 2 in the respective treatments. "Treatment dummies" comprises a set of dummy variables capturing treatment-specific period 1 satisfaction (resp. perceived difficulty). Male and Age are further individual-level controls. P-values of test of equality in coefficients are reported at the bottom of the table. For example, the line "Ho: EXRA x P2 = RANK x P2" provides the p- value of a test of equality between the "EXRA x P2" and the "RANK x P2" coefficients. Levels of significance: $*p < 0.1, **p < 0.05, ***p < 0.01$.

Table E.2: The effects of different randomly assigned reference workers

	Satisfaction		Difficulty	
	(1)	(2)	(3)	(4)
EXRA-HI x P2	-0.231***	-0.231***	0.661***	0.661***
	(0.053)	(0.053)	(0.050)	(0.050)
EXRA-MI x P2	-0.098*	-0.098*	0.600***	0.600***
	(0.052)	(0.052)	(0.046)	(0.046)
EXRA-LO x P2	0.082**	0.082**	0.292***	0.292***
	(0.041)	(0.041)	(0.046)	(0.046)
EXRA-NO x P2	-0.078	-0.078	0.416***	0.417***
	(0.049)	(0.049)	(0.045)	(0.045)
RANK x P2	-0.180***	-0.180***	0.467***	0.467***
	(0.035)	(0.035)	(0.032)	(0.032)
Male		-0.068**		0.013
		(0.033)		(0.046)
Age		0.004***		-0.000
		(0.001)		(0.002)
Treatment dummies	Yes	Yes	Yes	Yes
Ho: EXRA-HI x P2 = EXRA-MI x P2	0.074	0.073	0.367	0.367
Ho: EXRA-HI x P2 = EXRA-LO x P2	0.000	0.000	0.000	0.000
Ho: EXRA-HI x P2 = EXRA-NO x P2	0.034	0.034	0.000	0.000
Ho: EXRA-MI x P2 = EXRA-LO x P2	0.007	0.007	0.000	0.000
Ho: EXRA-MI x P2 = EXRA-NO x P2	0.776	0.778	0.004	0.004
Ho: EXRA-LO x P2 = EXRA-NO x P2	0.013	0.013	0.052	0.052
R^2	0.936	0.936	0.811	0.811
Observations	3044	3044	3044	3044

Note: OLS estimations. "EXRA-HI x P2" is a dummy which equals 1 if the participant was assigned to the EXRA-HI treatment. "EXRA-MI x P2" is a dummy which equals 1 if the participant was in the EXRA-MI treatment. These coefficients indicate by how much satisfaction (resp. perceived difficulty) changed from period 1 to period 2 in the respective treatments. The remaining interactions variables are defined in a similar way. "Treatment dummies" comprises a set of dummy variables capturing treatment-specific period 1 satisfaction (resp. perceived difficulty). Male and Age are further individual-level controls. P-values of test of equality in coefficients are reported at the bottom of the table. For example, the line "Ho: EXRA-HI x P2 = EXRA-MI x P2" provides the p-value of a test of equality between the "EXRA-HI x P2" and the "EXRA-MI x P2" coefficients. Levels of significance: $*p < 0.1, **p < 0.05, ***p < 0.01$.

Table E.3: The effects of different randomly assigned reference workers on satisfaction (heterogeneity)

	e1 < LOW		LOW < e1 < MED		MED < e1 < HI		e1 > HI	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
EXRA-HI x P2	-0.185	-0.187	-0.248***	-0.248***	-0.322***	-0.322***	0.176	0.176
	(0.154)	(0.155)	(0.081)	(0.081)	(0.082)	(0.083)	(0.181)	(0.181)
EXRA-MI x P2	-0.321*	-0.321*	-0.263***	-0.263***	-0.005	-0.005	0.548***	0.548***
	(0.179)	(0.179)	(0.089)	(0.089)	(0.072)	(0.072)	(0.119)	(0.119)
EXRA-LO x P2	0.125	0.125	0.045	0.044	0.053	0.053	0.333*	0.333*
	(0.161)	(0.161)	(0.063)	(0.063)	(0.057)	(0.057)	(0.177)	(0.177)
EXRA-NO x P2	0.159	0.159	-0.136*	-0.136*	-0.066	-0.066	-0.106	-0.106
	(0.159)	(0.159)	(0.074)	(0.074)	(0.079)	(0.079)	(0.178)	(0.178)
RANK x P2	0.141	0.141	-0.202***	-0.202***	-0.229***	-0.229***	-0.079	-0.079
	(0.138)	(0.138)	(0.053)	(0.053)	(0.051)	(0.051)	(0.172)	(0.172)
Male		-0.105		-0.007		-0.117**		-0.049
		(0.135)		(0.052)		(0.049)		(0.124)
Age		0.005		0.005***		0.005**		0.004
		(0.005)		(0.002)		(0.002)		(0.005)
Treatment dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Ho: EXRA-HI x P2 = EXRA-MI x P2	0.565	0.573	0.898	0.898	0.004	0.004	0.088	0.089
Ho: EXRA-HI x P2 = EXRA-LO x P2	0.165	0.164	0.005	0.005	0.000	0.000	0.536	0.537
Ho: EXRA-HI x P2 = EXRA-NO x P2	0.121	0.121	0.308	0.308	0.025	0.026	0.266	0.267
Ho: EXRA-MI x P2 = EXRA-LO x P2	0.064	0.065	0.005	0.005	0.535	0.535	0.316	0.317
Ho: EXRA-MI x P2 = EXRA-NO x P2	0.046	0.046	0.272	0.272	0.569	0.567	0.002	0.003
Ho: EXRA-LO x P2 = EXRA-NO x P2	0.880	0.880	0.064	0.065	0.226	0.224	0.081	0.082
R^2	0.904	0.905	0.940	0.940	0.939	0.939	0.939	0.939
Observations	255	255	1250	1250	1288	1288	251	251

Note: OLS estimations. "EXRA-HI x P2" is a dummy which equals 1 if the participant was assigned to the EXRA-HI treatment. "EXRA-MI x P2" is a dummy which equals 1 if the participant was in the EXRA-MI treatment. These coefficients indicate by how much satisfaction changed from period 1 to period 2 in the respective treatments. The remaining interactions variables are defined in a similar way. "Treatment dummies" comprises a set of dummy variables capturing treatment-specific period 1 satisfaction. Male and Age are further individual-level controls. P-values of test of equality in coefficients are reported at the bottom of the table. For example, the line "Ho: EXRA-HI x P2 = EXRA-MI x P2" provides the p-value of a test of equality between the "EXRA-HI x P2" and the "EXRA-MI x P2" coefficients. Sample divided into 4 subsamples: a) workers with a production in round 1 that is lower than the production in round 1 of the least productive reference worker ($e_1 \leq e_1(LO)$), b) workers with a production in round 1 that is between the production in round 1 of the least productive reference worker and the average reference worker ($e_1(LO) \leq e_1 \leq e_1(MI)$), c) workers with a production in round 1 that is between the production in round 1 of the average reference worker and the most productive reference worker ($e_1(MI) \leq e_1 \leq e_1(HI)$), d) workers with a production in round 1 that is higher than the most productive reference worker ($e_1(HI) \leq e_1$). Levels of significance: $^*p < 0.1, ^{**}p < 0.05, ^{***}p < 0.01$.

Table E.4: The effects of different randomly assigned reference workers on perceptions of task difficulty (heterogeneity)

	e1 < LOW		LOW < e1 < MED		MED < e1 < HI		e1 > HI	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
EXRA-HI x P2	0.835***	0.836***	0.610***	0.610***	0.688***	0.688***	0.647***	0.647***
	(0.198)	(0.198)	(0.081)	(0.081)	(0.074)	(0.074)	(0.142)	(0.143)
EXRA-MI x P2	0.268**	0.268**	0.742***	0.742***	0.615***	0.615***	0.262**	0.262**
	(0.126)	(0.127)	(0.076)	(0.076)	(0.070)	(0.070)	(0.118)	(0.118)
EXRA-LO x P2	0.225	0.225	0.275***	0.274***	0.301***	0.301***	0.375***	0.375***
	(0.180)	(0.181)	(0.063)	(0.063)	(0.078)	(0.078)	(0.139)	(0.139)
EXRA-NO x P2	0.364***	0.364***	0.477***	0.477***	0.425***	0.425***	0.170	0.170
	(0.122)	(0.122)	(0.068)	(0.068)	(0.073)	(0.073)	(0.150)	(0.150)
RANK x P2	0.269**	0.269**	0.477***	0.477***	0.479***	0.479***	0.556***	0.556***
	(0.117)	(0.117)	(0.050)	(0.050)	(0.047)	(0.047)	(0.124)	(0.124)
Male		0.028		-0.123*		0.049		0.092
		(0.156)		(0.074)		(0.068)		(0.166)
Age		-0.007		0.003		0.002		0.012*
		(0.005)		(0.003)		(0.003)		(0.006)
Treatment dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Ho: EXRA-HI x P2 = EXRA-MI x P2	0.016	0.016	0.236	0.236	0.476	0.476	0.039	0.039
Ho: EXRA-HI x P2 = EXRA-LO x P2	0.023	0.024	0.001	0.001	0.000	0.000	0.172	0.173
Ho: EXRA-HI x P2 = EXRA-NO x P2	0.043	0.043	0.211	0.212	0.012	0.012	0.022	0.022
Ho: EXRA-MI x P2 = EXRA-LO x P2	0.846	0.846	0.000	0.000	0.003	0.003	0.535	0.536
Ho: EXRA-MI x P2 = EXRA-NO x P2	0.586	0.587	0.010	0.010	0.061	0.061	0.632	0.632
Ho: EXRA-LO x P2 = EXRA-NO x P2	0.525	0.526	0.029	0.029	0.249	0.249	0.317	0.318
R^2	0.769	0.771	0.803	0.803	0.825	0.825	0.842	0.843
Observations	255	255	1250	1250	1288	1288	251	251

Note: OLS estimations. "EXRA-HI x P2" is a dummy which equals 1 if the participant was assigned to the EXRA-HI treatment. "EXRA-MI x P2" is a dummy which equals 1 if the participant was in the EXRA-MI treatment. These coefficients indicate by how much perceived difficulty changed from period 1 to period 2 in the respective treatments. The remaining interactions variables are defined in a similar way. "Treatment dummies" comprises a set of dummy variables capturing treatment-specific period 1 perceived difficulty. Male and Age are further individual-level controls. P-values of test of equality in coefficients are reported at the bottom of the table. For example, the line "Ho: EXRA-HI x P2 = EXRA-MI x P2" provides the p-value of a test of equality between the "EXRA-HI x P2" and the "EXRA-MI x P2" coefficients. Sample divided into 4 subsamples: a) workers with a production in round 1 that is lower than the production in round 1 of the least productive reference worker ($e_1 \leq e_1(LO)$), b) workers with a production in round 1 that is between the production in round 1 of the least productive reference worker and the average reference worker ($e_1(LO) \leq e_1 \leq e_1(MI)$), c) workers with a production in round 1 that is between the production in round 1 of the average reference worker and the most productive reference worker ($e_1(MI) \leq e_1 \leq e_1(HI)$), d) workers with a production in round 1 that is higher than the most productive reference worker ($e_1(HI) \leq e_1$). Levels of significance: $^*p < 0.1, ^{**}p < 0.05, ^{***}p < 0.01$.

Table E.5: The effects of endogenously chosen reference workers and of targeted exogenous matching

	Satisfaction		Difficulty	
	(1)	(2)	(3)	(4)
RANK x P2	-0.180***	-0.180***	0.467***	0.467***
	(0.035)	(0.035)	(0.032)	(0.032)
EXRA x P2	-0.081***	-0.082***	0.493***	0.493***
	(0.025)	(0.025)	(0.023)	(0.023)
ENDO x P2	-0.198***	-0.198***	0.654***	0.654***
	(0.036)	(0.036)	(0.031)	(0.031)
EXBE x P2	-0.272***	-0.272***	0.718***	0.718***
	(0.053)	(0.053)	(0.046)	(0.046)
Male		-0.062**		-0.033
		(0.027)		(0.037)
Age		0.004***		-0.001
		(0.001)		(0.001)
Treatment dummies	Yes	Yes	Yes	Yes
Ho: EXRA x P2 = RANK x P2	0.021	0.021	0.497	0.497
Ho: EXRA x P2 = ENDO x P2	0.008	0.008	0.000	0.000
Ho: EXRA x P2 = EXBE x P2	0.001	0.001	0.000	0.000
Ho: ENDO x P2 = EXBE x P2	0.247	0.247	0.256	0.256
R^2	0.935	0.935	0.814	0.814
Observations	4548	4548	4548	4548

Note: OLS estimations. "RANK x P2" is a dummy which equals 1 if the participant was assigned to the RANK treatment. "EXRA x P2" is a dummy which equals 1 if the participant was in the EXRA treatment. These coefficients indicate by how much satisfaction (resp. perceived difficulty) changed from period 1 to period 2 in the respective treatments. The remaining interactions variables are defined in a similar way. "Treatment dummies" comprises a set of dummy variables capturing treatment-specific period 1 satisfaction (resp. perceived difficulty). Male and Age are further individual-level controls. P-values of test of equality in coefficients are reported at the bottom of the table. For example, the line "Ho: EXRA x P2 = RANK x P2" provides the p-value of a test of equality between the "EXRA x P2" and the "RANK x P2" coefficients. Levels of significance: $*p < 0.1, **p < 0.05, ***p < 0.01$.

Table E.6: The effects of monetary incentives and social comparisons

	Satisfaction				Difficulty			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Piece-rate (pooled) x P2	-0.156***	-0.156***			0.621***	0.621***		
	(0.026)	(0.026)			(0.023)	(0.023)		
Flat wage (pooled) x P2	-0.206***	-0.206***			0.591***	0.591***		
	(0.023)	(0.023)			(0.020)	(0.020)		
RANK x P2			-0.180***	-0.180***			0.467***	0.467***
			(0.035)	(0.035)			(0.032)	(0.032)
RankxDOLLAR x P2			-0.100*	-0.100*			0.511***	0.511***
			(0.051)	(0.051)			(0.041)	(0.041)
ENDO x P2			-0.198***	-0.198***			0.654***	0.654***
			(0.036)	(0.037)			(0.031)	(0.031)
EndoxDOLLAR x P2			-0.150***	-0.150***			0.619***	0.619***
			(0.035)	(0.035)			(0.033)	(0.033)
EXBE x P2			-0.272***	-0.272***			0.718***	0.718***
			(0.053)	(0.053)			(0.046)	(0.046)
EXBExDOLLAR x P2			-0.224***	-0.224***			0.738***	0.738***
			(0.055)	(0.055)			(0.047)	(0.047)
Male		-0.090***		-0.089***		-0.042		-0.041
		(0.027)		(0.027)		(0.037)		(0.037)
Age		0.004***		0.004***		-0.002		-0.002
		(0.001)		(0.001)		(0.001)		(0.001)
Treatment dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Ho: Flat wage x P2 = Piece-rate x P2	0.148	0.148			0.327	0.326		
Ho: RANK$ x P2 = RANK x P2			0.197	0.197			0.394	0.394
Ho: RANK$ x P2 = ENDO x P2			0.121	0.121			0.006	0.006
Ho: RANK$ x P2 = EXBE x P2			0.020	0.020			0.001	0.001
Ho: RANK$ x P2 = EXBE$ x P2			0.099	0.099			0.000	0.000
Ho: RANK$ x P2 = ENDO$ x P2			0.421	0.422			0.043	0.043
Ho: EXBE$ x P2 = ENDO$ x P2			0.259	0.259			0.041	0.041
R2	0.933	0.934	0.934	0.934	0.811	0.811	0.811	0.811
Observations	4504	4504	4504	4504	4504	4504	4504	4504

Note: OLS estimations. "Piece-rate (pooled) x P2" is a dummy which equals 1 the treatment that offered a piece-rate in round 2. "Flat wage (pooled) x P2" is participant was in one of the treatment that did not offer a piece-rate in round 2. "RANK x P2" is a dummy which equals 1 if the participant was assigned to the RANK treatment. These coefficients indicate by how much satisfaction (resp. perceived difficulty) changed from period 1 to period 2 in the respective treatments. The remaining interactions variables are defined in a similar way. "Treatment dummies" comprises a set of dummy variables capturing treatment-specific period 1 satisfaction (resp. perceived difficulty). Male and Age are further individual-level controls. P-values of test of equality in coefficients are reported at the bottom of the table. For example, the line "Ho: RANK$ x P2 = RANK x P2" provides the p- value of a test of equality between the "RANK$ x P2" and the "RANK x P2" coefficients. Levels of significance: $*p < 0.1, **p < 0.05, ***p < 0.01$.

F Additional material related to the remaining questionnaire measures

F.1 Socio-demographics

- What is your gender? [male/female]
- In which year were you born? [1900-2010]
- What is your monthly gross income? [brackets]
- Which of the following best describes your race or ethnicity? [Caucasian / White, African American / Black, Hispanic/Latino, Asian American / Asian, Native American, Other]
- What category best describes your highest level of education? [8th grade or less, some high school, high school degree / GED, Some college, 2-year College Degree 4-year College Degree, Master's Degree, Doctoral Degree, Other]
- In which state do you currently reside? [list of states]
- Many people in the USA lean towards a political party. Which party do you lean towards? [Democrats, Republicans, Other, None]

F.2 Post-effort questions

After both Parts 1) and 2), we ask

- On a scale from 1-5, how difficult did you find the task? [1. Not at all difficult, ... ,5. Very difficult]
- On a scale from 1-5, how stressed have you been while completing the task? [1. Not at all stressed, ... ,5. Very stressed]
- How satisfied are you with your performance? [1. Not at all satisfied, ... ,5. Very satisfied]

F.3 Exit survey

To all participants who get to see a reference worker, we ask:

- Please describe in a few sentences how the performance of the other participant affected your performance (open-ended).

- On a scale from -5 to +5, how did observing the performance of the other participant affect your performance? [-5. Negatively affected my perf., ... ,0. Did not affect my perf., ... ,+5. Positively affected my perf.]
- On a scale from -5 to +5, did observing the performance of the other participant motivate you or discourage you? [-5. Discouraged me a lot, ... ,0. Did not affect me, ... ,5. Motivated me a lot]
- On a scale from 1 to 5, did observing the performance of the other participant make you nervous? [1. Not at all nervous, ... ,5. Very nervous]
- On a scale from 1 to 5, to what degree did you feel in competition with the other participant did you feel? [1. No competition at all, ... ,5. Very high competition]
- On a scale from 1 to 5, did observing the performance of the other participant make the task more enjoyable for you? [1. Not at all more enjoyable, ... ,5. Much more enjoyable]

In addition, we ask a set of "counterfactual questions" to assess how people think they would have performed, had they been assigned a different reference worker. In the EXO (and EXO-BEST) treatments, for example, we ask :

- In the previous round, you observed the performance of the reference participant who ranked 4th. Imagine that, instead of observing the reference participant who ranked 4th, you had been assigned the reference participant who was ranked 26. How would this have affected you? [A. It would have increased my performance, compared to the performance I achieved while observing the reference participant ranked 4th. B. It would have decreased my performance, compared to the performance I achieved while observing the reference participant ranked 4th. C. It would have made no difference.]
- Imagine that, instead of observing the reference participant who ranked 4th, you had been assigned the reference participant who was ranked 49. How would this have affected you? [A. It would have increased my performance, compared to the performance I achieved while observing the reference participant ranked 4th. B. It would have decreased my performance, compared to the performance I achieved while observing the reference participant ranked 4th. C. It would have made no difference.]
- Finally, imagine that instead of observing the reference participant who ranked 3rd, you had been assigned NO reference participant. How would this have affected you? [A. It would have increased my performance, compared to the performance I achieved while observing the reference participant ranked 4th; B. It would have decreased my performance, compared to the performance I achieved while observing the reference participant ranked 4th; C. It would have made no difference.]
- Could you have chosen a reference participant, which reference participant would you have chosen? [Participant ranked 4, participant ranked 26, participant ranked 49, None]

In the ENDO treatment, we ask

- In the previous round, you observed the performance of the reference participant who ranked XXXth. Please indicate in a few sentences why you have chosen to observe the performance of this reference participant. (Open answer)
- Please describe in a few sentences how the performance of the other participant affected your performance. (Open answer)
- On a scale from 1-5, do you regret to have chosen this reference participant? [1. Not regrets at all, ... ,5. A lot of regrets]

Finally, in the EXO-NO RP we ask the following counterfactual questions:

- In the previous round, you could not observe the performance of a reference participant. Imagine that you had been assigned the reference participant who was ranked 4th. How would this have affected you? [A. It would have increased my performance, compared to not observing a reference participant. B. It would have decreased my performance, compared to not observing a reference participant. C. It would have made no difference.]
- Imagine that you had been assigned the reference participant who was ranked 26th. How would this have affected you? [A. It would have increased my performance, compared to not observing a reference participant. [B. It would have decreased my performance, compared to not observing a reference participant. C. It would have made no difference.]
- Finally, imagine that you had been assigned the reference participant who was ranked 59. How would this have affected you? [A. It would have increased my performance, compared to not observing a reference participant. B. It would have decreased my performance, compared to not observing a reference participant. C. It would have made no difference.]

while in the ENDO treatment, if a subject decided to see no reference worker we ask:

- In the previous round, you decided not to observe a reference participant. Please indicate in a few sentences why you made this choice.(open answer)
- On a scale from 1-5, do you regret to have chosen not to observe a reference participant? [1. Not regrets at all, ... ,5. A lot of regrets]