Senn, Julien; Schmitz, Jan; Zehnder, Christian

Working Paper

Leveraging Social Comparisons: The Role of Peer Assignment Policies for Productivity and Stress

CESifo Working Paper, No. 11972

Provided in Cooperation with:
Ifo Institute – Leibniz Institute for Economic Research at the University of Munich

Suggested Citation: Senn, Julien; Schmitz, Jan; Zehnder, Christian (2025) : Leveraging Social Comparisons: The Role of Peer Assignment Policies for Productivity and Stress, CESifo Working Paper, No. 11972, CESifo GmbH, Munich

This Version is available at:
https://hdl.handle.net/10419/324963

CESifo WORKING PAPERS

11972
2025

July 2025

Leveraging Social Comparisons: The Role of Peer Assignment Policies for Productivity and Stress

Julien Senn, Jan Schmitz, Christian Zehnder

CESifo

Impressum:

CESifo Working Papers
ISSN 2364-1428 (electronic version)
Publisher and distributor: Munich Society for the Promotion of Economic Research - CESifo GmbH
The international platform of Ludwigs-Maximilians University's Center for Economic Studies and the ifo Institute
Poschingerstr. 5, 81679 Munich, Germany
Telephone +49 (0)89 2180-2740, Telefax +49 (0)89 2180-17845, email office@cesifo.de
Editor: Clemens Fuest
https://www.ifo.de/en/cesifo/publications/cesifo-working-papers
An electronic version of the paper may be downloaded
· from the SSRN website: www.SSRN.com
· from the RePEc website: www.RePEc.org
· from the CESifo website: https://www.ifo.de/en/cesifo/publications/cesifo-working-papers

Leveraging Social Comparisons: The Role of Peer Assignment Policies for Productivity and Stress*

Julien Senn Jan Schmitz Christian Zehnder

July 2, 2025

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 an increase in pay of about 10 percent. A downside of targeted peer assignment is that, unlike endogenous peer selection, it leads to a large increase in stress. 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 show 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

*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., Villeval, 2020; Falk and Ichino, 2006). While social comparisons often arise spontaneously, an increasing number of firms purposefully try to make them more salient—e.g., through dynamic, computerized, leaderboards. But can social comparisons really be leveraged to boost productivity, and can this have negative, unintended consequences such as, e.g., affecting workers' psychological well-being?

To date, the literature has mostly 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 be possible to further leverage social comparisons by relying on alternative peer assignment mechanisms.

A simple alternative to random assignment is to let workers choose whom to compare to. Such self-chosen comparisons are pervasive in many contexts (see, e.g., Fujita and Diener, 1997; Suls et al., 2002; Kiessling et al., 2021), including the workplace.[1] If workers know what type of comparison motivates them most, endogenous peer choice might be highly effective. 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 unmotivating peers (e.g., by comparing downwards to feel good about themselves).

Another potentially promising way to improve productivity is to exogenously assign workers to peers that are predicted to be highly motivating (see, e.g., Graham et al., 2014; Roels and Su, 2014; Kräkel, 2016; Carrell et al., 2013; Chen and Gong, 2018). Practical attempts that go in this direction can be seen in the recent trend to "gamify" the provision of information about coworkers' productivity to boost output.[2] However, systematic exogenous assignment procedures might backfire if workers get upset about being pressured to observe peers they would *not* have chosen. Moreover,

[1] For example, researchers compare output to selected colleagues, wealth managers compare portfolio returns to particular competitors, and schoolchildren compare grades with some classmates.

[2] This practice is increasingly adopted across industries—from sales to (online) retail and banking. Examples include Target, Amazon, and Disney (see, e.g., https://tinyurl.com/ycystmez). For more on the rise of gamification and further examples, see Koivisto and Hamari (2019).

implementing such systems requires detailed information and might entail considerable costs.

Surprisingly, evidence on the effects of such policies for performance remains scarce. The psychological consequences of social comparisons have also largely been ignored by the literature cited above. However, recent evidence from psychology suggests that social comparisons might affect stress and psychological well-being (e.g., Reiff et al., 2022; Buunk and Dijkstra, 2017; Bárcena-Martín et al., 2017; Fujita and Diener, 1997). Hence, stress might be a particularly important dimension to consider given that it is associated with a number of (negative) labor market outcomes such as lower productivity (Halkos and Bousinakis, 2010), higher absenteeism (Jacobson et al., 1996; Leontaridi and Ward-Warmedinger, 2002), and higher turnover intentions (Mosadeghrad, 2013). It is also associated with poor mental health and depression (Bianchi et al., 2017), which have been shown to hamper human capital accumulation (Currie and Stabile, 2006; Fletcher, 2010; Eisenberg et al., 2009) and to undermine important labor market outcomes such as earnings and employment (e.g., Bubonya et al., 2017; Ettner et al., 1997; Stewart et al., 2003; Fletcher, 2014; Biasi et al., 2021).

This paper sheds light on the effects of different peer assignment mechanisms for productivity and stress. Our large-scale (N=6532), pre-registered real-effort experiment provides new causal evidence on the 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. Participants are hired to perform a simple real-effort task over two consecutive periods. The first period is identical across all treatments: workers complete the task in isolation. Each worker then receives private information about their performance relative to "*60 other participants who already took part in this study*" (the "reference population"). At the start of the second period, workers are randomized into different conditions that vary whether and how they are matched with a peer (a "reference worker").

The same three reference workers are drawn from the reference population in all treatments: a high-, average-, and low-productivity worker. Social comparisons are operationalized by providing participants with *real-time* information about the second-period performance of the reference worker. All treatments also include the option of not observing a reference worker.

In the EXRA (Exogenous Random) treatment, the assignment to a reference worker is random. In the EXBE (Exogenous Best) treatment, participants are exogenously assigned to the reference worker predicted to have the largest positive impact on their performance. This prediction is based on data from EXRA and corresponds to the high-productivity reference worker for nearly all participants. In the ENDO (Endogenous) treatment, participants *choose* their preferred reference worker or no reference worker. Finally, in the RANK treatment, reference workers are not mentioned—participants receive *only* information about how they rank relative to the reference population. These treatments allow us to cleanly identify the causal effects of social comparisons under different peer assignment mechanisms. To test the robustness of our results and benchmark their magnitude, we also introduce treatments that vary the period 2 compensation scheme (fixed wage vs. performance pay).

Our setting resembles work environments where workers can observe each other and 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* the output of a peer, controlling for other confounding factors. We therefore consciously restrict our attention to a setting that neither involves production complementarities between workers, nor provides scope for social learning. Importantly, workers in our experiment may only observe the *productivity* of a peer, but they can never observe any other peer characteristic. This design feature is important because the effects of peer productivity could be confounded by other characteristics if workers were randomly assigned to peers that also vary in other dimensions that are not orthogonal to productivity. In addition, our design circumvents multiple hurdles that often complicate the identification of social spillovers, such as the reflection problem (Manski, 1993) and informational confounds (Charness et al., 2013). We elaborate on these points in Section 3.3.

We document several novel findings. First, when assessing the causal effects of comparisons with randomly assigned peers, we show that social comparisons entail an important trade-off: participants exposed to a more productive reference worker not only become more productive, but also report a stronger increase in stress. For example, participants who are randomly assigned to the least productive peer experience an increase in productivity of 10.6% and an increase in stress of 26% of

a standard deviation. In contrast, those who are assigned to the most productive peer experience a significantly larger increase in output (+14.2%) and a much larger increase in stress (+55% of a standard deviation). Thus, social comparisons have motivational potential, but they also entail non-negligible psychological costs.

Second, we investigate how the effects of social comparisons differ depending on whether peers are endogenously chosen (ENDO) or exogenously assigned to maximize expected productivity (EXBE). Participants in ENDO exhibit a productivity increase that is 18% larger than in EXRA, but *not* significantly different than in EXBE. In terms of magnitude, the observed effect size in ENDO and EXBE is comparable to that of introducing an incentive that raises pay by about 10%. While both ENDO and EXBE yield very similar productivity effects, they have very different implications for stress: EXBE yields an increase in stress that is 15% larger than in ENDO. These results highlight the power of endogenous comparisons: allowing workers to choose whom to compare to significantly boosts productivity *without* increasing stress as much as assigning them to the predicted most motivating reference worker.

Next, we investigate why productivity gains in ENDO and EXBE are virtually indistinguishable, while the impact on perceived stress is different. We first explore workers' preferences for peers. We show that while almost all workers in EXBE are assigned to the high-performance reference worker, only about 45% of workers in ENDO choose to compare to this peer. The second most frequent choice is not to compare to anyone (30%). Text analysis on choice motives reveals that workers who aim at productivity improvements 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 productive *and most stressful* peer. However, a crucial questions remains: If fewer individuals choose to compare to the predicted most motivating peer, why is productivity in EXBE not substantially higher than in ENDO?

We find that the motivational potential of exogenously assigned peers is fully realized only when the assigned peer aligns with the worker's preference. When there is a mismatch between the preferred and assigned peer, the motivational effect diminishes—a phenomenon we term the "*mismatch effect*." While such mismatches are widespread in EXBE, they are absent by design in ENDO. Interestingly, mismatches

affect productivity but not stress: participants assigned to the most productive peer against their will do not exhibit a notable increase in performance, but still experience a substantial rise in stress. These findings help explain why EXBE generates a sharp increase in stress but no notable productivity gain relative to ENDO.

We quantify the role of a change in the composition in the reference workers (composition effect) and of the mismatch effect in explaining treatment differences between ENDO and EXBE using Gelbach decompositions (Gelbach, 2016). The composition effect leads to an increase in performance in EXBE (relative to ENDO), but this performance increase is more than offset by the large mismatch effect in EXBE. Turning to the difference in stress between ENDO and EXBE, we find that the composition effect is the largest contributor, an effect that is *not* offset because there is no mismatch effect in the stress dimension. In both the performance and the stress domain, the residual variation is small and negative, suggesting that—if anything—the mere act of choosing only marginally influences treatment differences.

Finally, we benchmark the effects of social comparisons using three treatments that interact social comparisons with financial incentives. Our findings are largely robust to the introduction of performance-based bonus payments, and we show that social comparisons and monetary incentives act as complements in our context. Moreover, a follow-up experiment aimed at contrasting the effects of social and non-social comparisons reveals that social comparisons generate much larger behavioral effects than comparable non-social goals. In fact, the performance increase from social comparisons is more than twice as large as that from comparable non-social goals. Moreover, social comparisons make workers substantially more stressed and more nervous than non-social goals.

Overall, our results highlight that social comparisons can, in principle, be leveraged to boost productivity—but policies aimed at increasing output may also carry unintended consequences, such as elevated worker stress. While these side effects are often overlooked, we argue that they warrant systematic monitoring, as they may ultimately impact firms' overall performance.

2 Related Literature

Our paper relates to multiple strands of the literature and makes several contributions. First, our paper relates to the literature on the effects of social comparisons on productivity. Existing studies have focused on the effects of relative performance feedback (see, e.g., Charness et al., 2013; Gill et al., 2019; Azmat and Iriberri, 2016; Eriksson et al., 2009; Drouvelis and Paiardini, 2022; Kuhnen and Tymula, 2012) and randomly assigned peers on productivity (see, e.g., Falk and Ichino, 2006; Bellemare et al., 2010; Rosaz et al., 2016; Mas and Moretti, 2009; Bandiera et al., 2010).[3] We contribute to this literature by highlighting the central role of peer assignment mechanisms for motivational spillovers and by establishing the effects of social comparisons for stress—a dimension that has been largely ignored in this literature.

Moreover, we improve upon existing work by distinguishing the effects of social comparisons with those of non-social goals (see, e.g., Corgnet et al., 2015, 2018)[4] and by cleanly isolating the behavioral effects arising from observing peer productivity, 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 make a methodological contribution by developing an experimental paradigm that permits the identification of the social spillovers of *observing* peers while circumventing the main hurdles pertaining their estimation (Manski, 1993). While we make these contributions in a stylized framework, we have reasons to believe that our results generalize to other settings: A recent meta-study on peer effects shows that the *magnitude* of productivity spillovers from one worker to another is very similar in the lab and the field (Herbst and Mas, 2015). More generally, there is now broad agreement that *qualitative* results from laboratory studies are generally externally valid (Kessler and Vesterlund, 2015; Horton et al., 2011).

Our paper also relates to the growing literature interested in non-random peer assignment mechanisms. Recent papers have theorized that exogenous peer-

[3]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).

[4]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).

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 work context and to contrast it with alternative peer assignment mechanisms. We are also not aware of any study on the effects of endogenously chosen peers for work performance *and stress*. In a related study, Kiessling et al. (2021) looked at the performance effects of self-selected peers in the context of a running contest organized at school.[5] Our study differs from theirs in many ways. Most importantly, their setup i) involves simultaneous interactions, thereby raising the question of a possible reflection problem (Manski, 1993), ii) does not control for potential feedback effects, and iii) may allow peer characteristics beyond productivity to entangle motivational spillovers with other peer effects. In contrast, our experimental paradigm cleanly accounts for all these identification challenges. Moreover, we implement a targeted matching policy (EXBE), we assess the effects of social comparisons for stress, we establish whether and how social comparisons and financial incentives interact, and we compare social and non-social goals.

Our paper also connects to the growing literature on mental health in economics (e.g., Cobb-Clark et al., 2022; De Quidt and Haushofer, 2016; Ridley et al., 2020; Roth et al., 2024a,b). In particular, it relates to the recent work that has explored the effects of peers for mental health among adolescents (Bütikofer et al., 2023; Kiessling and Norris, 2023; Braghieri et al., 2022). We contribute to this literature by showing that stress increases in the productivity of the observed peer. We also show that letting workers choose whom to compare to mitigates the increase in stress, and we show that social comparisons generate significantly more stress and more anxiety than comparable, non-social, goals. More generally, our paper connects to the literature linking stress and mental health with (negative) labor market outcomes (see, e.g., Bubonya et al., 2017; Fletcher, 2014, 2010; Biasi et al., 2021; Stewart et al., 2003; Leontaridi and Ward-Warmedinger, 2002; Mosadeghrad, 2013; Ettner et al., 1997).

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

3 Experimental Design

Figure 1 provides an overview of the experimental design and of the main treatments.[6] 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. 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 received real-time feedback about their own output. As we explain below, these participants constitute 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. These participants also started with a first round of effort provision during which they received real-time feedback about their own output. Upon completion of this first round, they *privately* learned how their productivity in round 1 compares with the productivity in round 1 of the workers from the reference population ('Feedback'). They were 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 a second time.

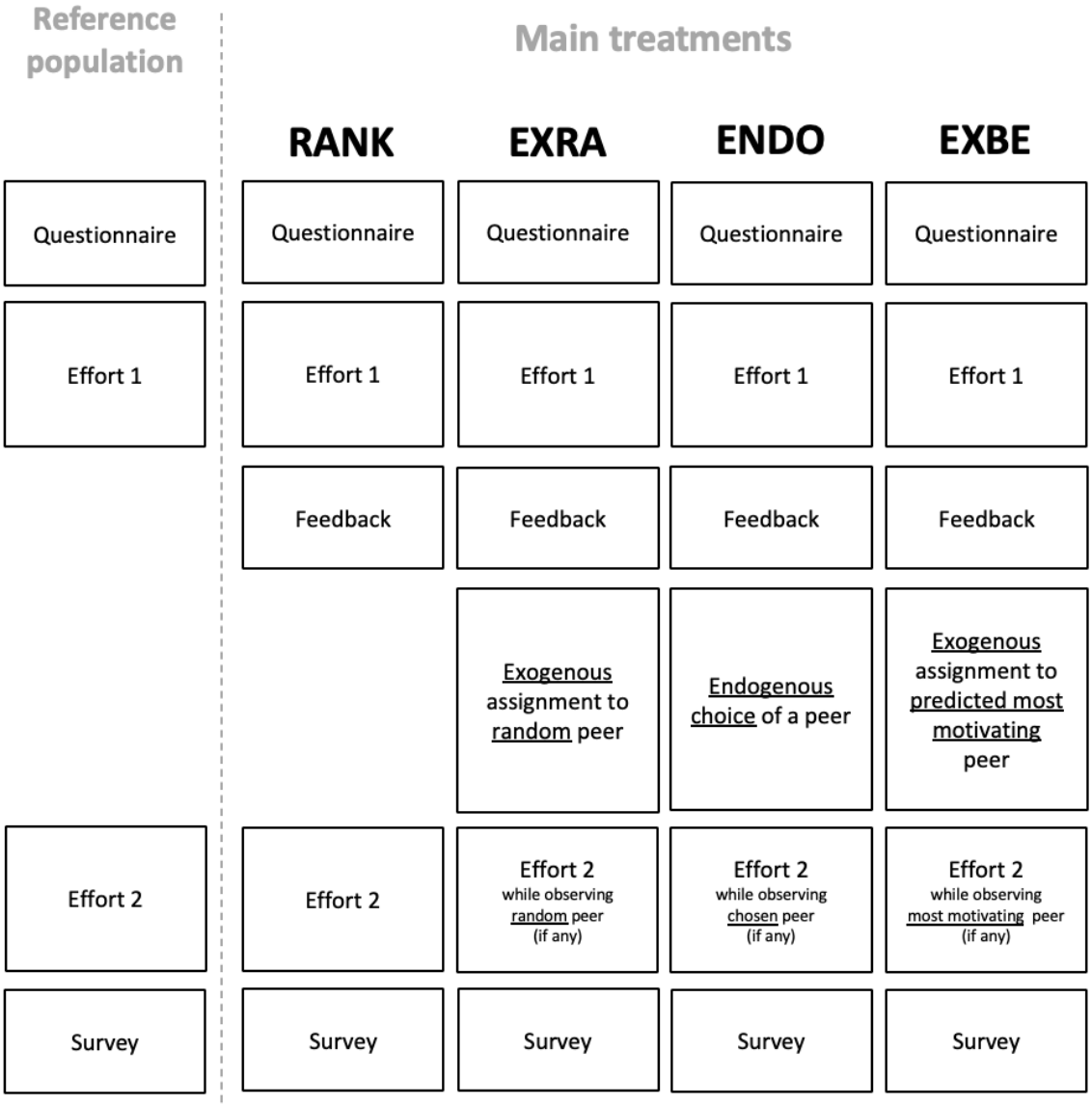
In the following, we provide details on the real-effort task, the different treatments, the reference population and the reference workers, as well as the sample.

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.

[6] An overview of the entire experimental design is provided in Figure A.4 in Appendix.

Figure 1: Overview of the main experimental design



Note: Our experimental design comprises two set of participants. The reference population (left most column), which we use as a source of (social) information to be provided to the main participants. The main participants are randomized (between-subjects) into different treatments that vary *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, and *how* participants are matched with a reference worker. Here, we only summarize our main treatments in which participants are paid a flat wage. In Figure A.4 in Appendix, we provide a detailed overview of the entire study.

The so-called a-b-task fulfills all the above mentioned requirements. It 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. 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.

This task shares the main characteristics of typical clerical and manual jobs and has been widely used to study worker motivation (DellaVigna and Pope, 2017; Butera et al., 2022; Berger and Pope, 2011; Amir and Ariely, 2008). Most importantly, it is effort-intensive, repetitive, and tiring. These features also characterize the simple jobs that are typically studied in field studies on worker motivation, such as data-entry (e.g., Kube et al., 2012) or fruit-picking (see, e.g., Bandiera et al., 2010).

3.2 Exogenous Assignment to Random Reference Workers (EXRA)

We first describe the details of the treatment in which workers are exogenously assigned to a random reference workers (EXRA).[7] 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.[8] They were made aware that the study would consist of several parts, and that they would receive instructions separately 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. Period-1 performance is therefore fully comparable across treatments.

Part 1: Questionnaire The experiment started with questions on participants' socio-demographics (see Appendix G.2 for details).

[7] The relevant screenshots are provided in Appendix G.1.

[8] 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 robustness purposes, we also ran treatments in which workers were paid a piece rate (more details in section 4.4).

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” and ”b” buttons as quickly as possible during 5 minutes. Participants went through a practice round of 15 seconds. They were then asked to give an estimate of how many points they thought they would be able to reach, and were asked to work on the task for 5 minutes. While working on the task, participants were constantly updated on their current output and the remaining time (see the screenshot provided in Figure G.1 in the Appendix).

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 information 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 (see screenshot G.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 the 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 G.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 G.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.

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 G.4 for details). They were then informed about their profit and the payment procedure.

Stress elicitation We measured participants stress upon completion of each production period. 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 ("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."). We also asked whether they were satisfied with their performance, and whether they found the task difficult (see Appendix G.3 for details). An advantage of using a single-item question to measure stress is that it permits to measure it in a more obfuscated way (by "hiding it" around unrelated questions) than using a battery of questions.

Preferred reference worker We also elicited subjects' *preferred* reference worker, i.e. we asked each subject 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.

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, and only received feedback about their own performance. However, they 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.

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: 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).[9]

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 circumventing two problems that often plague (observational) studies on social comparisons. First, the reference population avoids the reflection problem (Manski, 1993) that arises when trying to identify social spillovers in the context of simultaneous interactions. 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 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 as a response to being observed.[10]

[9] In Appendix A.1, we provide the full distribution of the performance of our reference population in both periods. We also discuss the selection criteria for the three reference workers, and we depict their production paths in both rounds (Figures A.1 to A.3).

[10] If workers' mutual influence on each other were constant across treatments, the reflection problem would not threaten the identification of treatment effects. However, the different peer assignment

Another hurdle in the identification of social spillovers is that social comparisons also convey information about relative performance (we refer to this phenomenon as the "feedback problem"). Because relative 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 this type of feedback effect 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—thereby allowing us to cleanly isolate motivational spillovers and distinguish them from the effects of relative performance feedback.[11]

3.4 Description of the other treatments

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 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.

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 G.5 in Appendix G.1.1). Participants could also decide *not* to observe any reference worker.

mechanisms might affect how workers respond to observing each other. Our design guarantees a clean identification of motivational spillovers even in such a case where the reflection problem interacts with the treatments.

[11] Providing workers with rank information does not rule out all information effects. Individuals in treatments where they potentially compare themselves with a reference worker mechanically obtain information that comes on top of the rank information, but this additional information is part of the treatment.

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. Importantly, 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 G.3 in the Appendix). We provide more details on the tailoring of this matching procedure in Section 4.2.

In addition, we implemented three treatments (RANK$, ENDO$, EXBE$) in which participants were compensated for performance on top of their fixed payment. We also conducted a separate experiment to distinguish the effects of social comparisons from the effects of non-social comparisons. We will provide details on the implementation 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.[12]

3.5 Sample and experimental protocol

We pre-registered our study on the AEA RCT Registry (AEARCTR-0003217).[13] We ran our experiment on Amazon Mechanical Turk (MTurk)[14] in mid-August 2018. The

[12]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. 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.

[13]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 supplementary material.

[14]Because Mturk allows to assign a large set of small tasks to a very large amount of workers, it is no surprise that it is being increasingly used by academics, including economists, to conduct large-scale between-subjects studies (see e.g. DellaVigna and Pope, 2017; Doerrenberg et al., 2024; De Quidt et al., 2018; Almås et al., 2020; Cappelen et al., 2021, 2023). For example, DellaVigna and Pope (2017) also

experiment took about 10 minutes to complete, for which we paid a fixed wage of \$1.5. We provide more details on MTurk as well as on our recruitment protocol and exclusion criteria in Appendix A.3.

Our final sample comprises 6532 subjects.[15] We display the main summary statistics of our sample in Table A.3 in the Appendix. 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). 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. Last, we benchmark the effects of social comparisons and establish their robustness using treatments that interact social information with monetary incentives for production (RANK×\$, ENDO×\$, EXBE×\$). We also contrast the effects of social comparisons with those of non-social goals.

use MTurk and the a-b task to investigate the effects of different financial and non financial incentive schemes for worker motivation. 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, inattention, 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.

[15] 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.2.

4.1 The causal effects of randomly assigned reference workers

Figure 2a displays the average performance increase for each of the four sub-conditions of EXRA.[16] A very clear pattern emerges from the figure: On average, the performance gains from period 1 to period 2 increase systematically with the productivity of the assigned reference worker. Participants paired with the least productive reference worker (EXRA-LO) improve by 111 units (from 1051.2 to 1163.1; i.e., +10.6%, $p < 0.01$).[17,18] Those matched with the average-productivity peer (EXRA-MI) improve by 125 units (from 1021.2 to 1146.5; i.e., +12.3%, $p < 0.01$). The largest productivity gains, 148 units (from 1044.8 to 1192.8; i.e., +14.2%, $p < 0.01$), accrue for participants assigned to the most productive reference worker (EXRA-HI)—an effect that is significantly larger than the increase in performance induced by the low-productivity peer ($p = 0.03$). The difference between EXRA-HI and EXRA-MI, and the difference between EXRA-MI and EXRA-LO, are smaller in magnitude and not statistically significant ($p = 0.21$ and $p = 0.45$, respectively).

When comparing themselves to a reference worker, participants naturally gather information about their (relative) productivity. This information might in itself affect participants (see, e.g., Charness et al., 2013). In order to properly disentangle motivational spillovers from such feedback effects, it is instructive to compare participants in EXRA with those in RANK and in EXRA-NO. We find that participants who were not assigned to any reference worker in the EXRA condition (EXRA-NO) improve their performance by only 84 units (from 1043 to 1127.1; i.e., +8%, $p < 0.01$). This change in performance is not significantly different from the change in performance

[16] Workers in EXRA are randomly (and uniformly) assigned to one of the four treatment arms: no reference worker (EXRA-NO), the low productivity reference worker (EXRA-LO), the average productivity reference worker (EXRA-MI), or the high productivity reference worker (EXRA-HI). We therefore have approximately 500 observations per treatment arm (see Appendix A.2 for details).

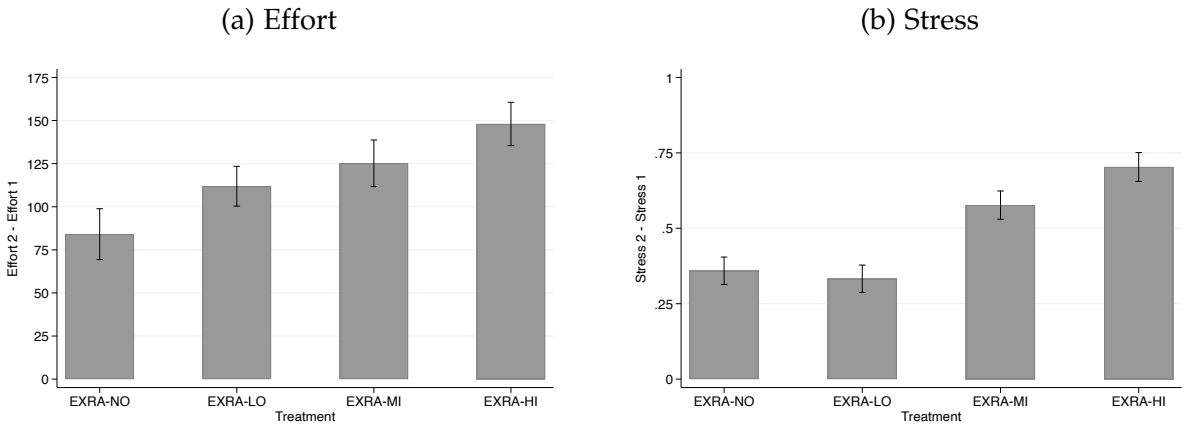
[17] All the p-values reported in this paper are based on Wald tests of linear hypotheses about the parameters of OLS estimations in which we regress the dependent variable on treatment dummies and interactions between treatment dummies and an indicator variable for period 2 (since treatments are implemented in period 2). An advantage of this procedure is that it also allows to control for a workers' individual characteristics. Without controls, the p-values 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 B.2.1. The p-values discussed in this section are taken from Table B.1.

[18] 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 ($p = 0.40$).

in the RANK treatment (p = 0.33) in which performance increases from 1040.6 to 1107.6 (+6.5%, $p < 0.01$).[19] The performance increases in EXRA-HI, EXRA-MI, and EXRA-LO, in contrast, are all significantly larger than in RANK (all three tests are highly significant, with $p < 0.01$).[20]

Together, these results show that social comparisons generate motivational spillovers that largely exceed those obtained by the sole provision of rank information. They also show that getting assigned to a more productive reference worker generates, on average, a higher increase in performance.

Figure 2: 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 +/- 1 standard error.

While we have demonstrated that randomly assigned reference workers generate substantial motivational spillovers, Figure 2b indicates that social comparisons also tend to generate a substantial increase in stress, particularly when participants are matched to productive peers.[21] Workers in EXRA-HI report a large increase in stress

[19]Note that not being assigned to a reference worker in EXRA is not exactly the same as being in the RANK condition because participants in RANK are not aware that other participants are assigned a reference worker.

[20]For completeness, note that the performance increase in EXRA-NO is significantly lower than in EXRA-HI ($p < 0.01$) and EXRA-MI ($p < 0.05$), but not significantly different than in EXRA-LO (p = 0.13).

[21]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. We also collected data on satisfaction about own performance and perceived task difficulty. As we show

of +0.70 (+55% of a standard deviation, $p < 0.01$), which is significantly different than the increase reported in all other treatment arms of EXRA.[22] Participants assigned to EXRA-MI report an increase in stress of +0.58 (+46% of a standard deviation), which is a significantly larger than the increase in stress in EXRA-LO ($p < 0.01$). Subjects assigned to no reference worker or to the least productive reference worker, in contrast, experience substantially lower increases in stress (+0.36 in EXRA-NO and +0.33 in EXRA-LO, i.e., +28% and +26% of a standard deviation, respectively, both $p < 0.01$).

In the RANK condition, perceived stress increases by +0.47 (+37% of a standard deviation, $p < 0.01$)–significantly less than when comparing to the average or the highly productive reference worker (both tests of equality in coefficients are significantly different from zero). Just like for the case of performance, the increase in stress in EXRA-NO is not significantly different from the increase documented in RANK. Together, these results provide causal evidence that observing a more productive reference worker not only increases productivity, but also substantially increases stress.[23]

In Appendix B.2.2, we explore whether the effects of the different reference workers depend on the characteristics of the observer. We find that virtually all workers experience the largest increase in productivity when assigned to the most productive reference worker (HI). 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 period 1. As we will discuss in Section 4.2, we base our targeted exogenous matching treatment (EXBE), which assigns workers to their predicted most motivating peer, on these results.

These first results highlight the potentially large effects that social comparisons

in Appendix F, the results are entirely consistent with the effects that social comparisons have on performance and on perceived stress.

[22]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-LO ($p < 0.01$) and EXRA-NO ($p < 0.01$).

[23] 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.

can have. Even in situations in which reference workers are *randomly* assigned, they can generate large increases in productivity. However, these increases in productivity are often accompanied by a substantial increase in stress. Importantly, these effects *depend on the productivity of the assigned reference worker*. These findings point to the relevance of the matching procedure and the conjecture that random assignment of peers most likely does *not* fully exploit the motivational potential of social comparisons. In the next section, we investigate the effects of two alternative assignment mechanisms that might further enhance social spillovers.

4.2 Leveraging social comparisons using non-random peer assignment policies

The first policy that we consider is EXBE. This condition exogenously assigns each worker to their predicted most motivating reference worker, based on their round 1 productivity and gender. To predict which reference worker is the most motivating for a particular worker, we use the data from the 2000 workers in EXRA to obtain a point estimate of expected performance for each possible 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. For the majority of participants, the predicted most motivating reference worker is the most productive one (HI), as discussed in the previous section. For details on the implementation of this assignment procedure, see Appendix B.6.

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 increasing perceived stress substantially because for most workers the high-productivity reference worker is predicted to be most motivating, but is also the most stressful peer to observe.

The second policy that we consider is ENDO, where participants are given the possibility to decide which reference worker to compare to (if any) in the second period. 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 (choice effect). The potential downside is 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.

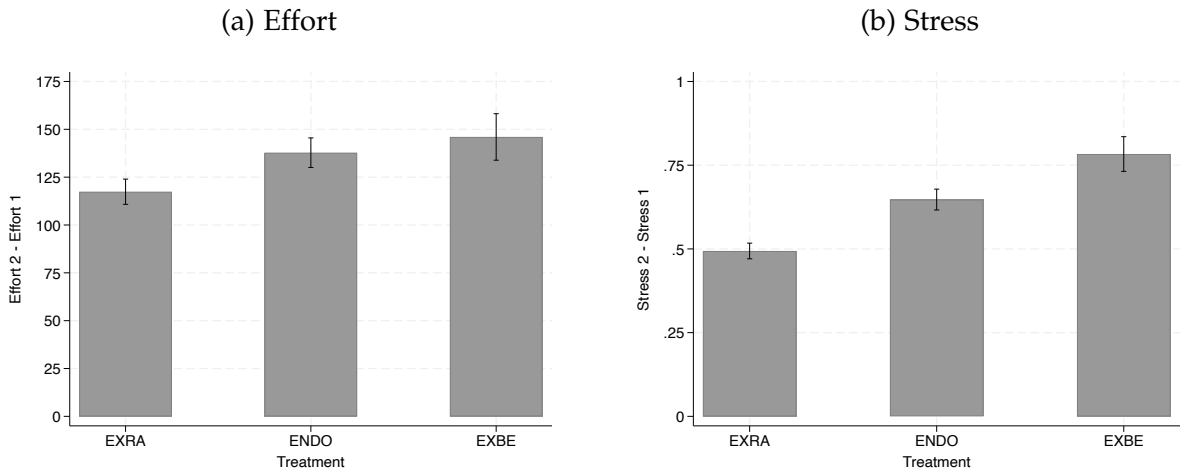
We depict the average change in productivity for EXBE and ENDO in Figure 3a below, along with the average effects of randomly assigned reference workers (EXRA). The figure unambiguously shows that both EXBE and ENDO generate productivity increases that are larger than the one documented in EXRA: while participants in EXRA improve by an average of 117 units (from 1039.9 in period 1 to 1157.3 in period 2, i.e., +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., +14.2 percent, $p < 0.01$) and participants in ENDO improve by an average of 138 units (from 1059.3 in period 1 to 1197.12 in period 2, i.e., +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 3b shows that these two treatments affect workers' stress levels differently. Forcing workers to compare to their predicted most motivating peer yields a significantly larger increase in stress (+0.78 points, +61% of a standard deviation, $p < 0.01$) than letting workers choose whom to compare to (+0.65 points, +53% of a standard deviation, $p < 0.01$; EXBE vs. ENDO: $p = 0.02$).[24]

Because performance increases slightly more in EXBE than in ENDO (although *not* significantly), one might argue that the lower increase in stress in ENDO comes at a cost. To capture this potential trade-off, we constructed a measure of "stress-adjusted" output by dividing performance in period 2 by stress in period 2. As we show in Appendix B.4, stress-adjusted output is significantly higher in ENDO than in EXBE ($p < 0.01$). This implies that participants generate more output per unit

[24]Compared to EXRA (+0.49 points, +38% of a standard deviation, $p < 0.01$), both EXBE and ENDO create larger increases in stress (ENDO vs. EXRA: $p < 0.01$; EXBE vs. EXRA: $p < 0.01$).

Figure 3: 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 +/- 1 standard error.

of stress in ENDO than in EXBE, i.e., that output increases faster than stress. In addition, we also show that ENDO generates a larger stress-adjusted output than EXRA (p = 0.037), and that EXBE generates a lower stress-adjusted output than EXRA (p = 0.049). These results are robust to defining stress-adjusted output as effort2/stress2 – effort1/stress1. There too, stress-adjusted output is higher in ENDO than in EXBE (although the statistical significance of the result is weaker, p = 0.08).

The comparison of EXBE with ENDO highlights the power of endogenous comparisons: letting workers choose whom to compare to generates a strong increase 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 responses 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.

4.3 Understanding the differences between endogenous choice and targeted matching

Why does ENDO produce a performance improvement similar in magnitude to EXBE, yet without a correspondingly large increase in perceived stress? We answer this question by exploring the role of three different channels through which endogenous choice might affect outcomes. First, we analyze the selection of reference workers in ENDO because choice patterns and their associated choice motives might have important implications for participants' performance and their perceived stress in period 2 ("*composition effect*"). Second, we study whether outcomes differ depending on whether participants observe their preferred reference worker or not. While participants' in ENDO always get to observe their preferred peer, this is not the case in EXBE where participants are exogenously assigned to reference workers and where a substantial proportion of participants may be matched with a reference worker they would *not* have chosen. These mismatches might have a substantial effect on productivity and stress ("*mismatch effect*"). Third, we will also consider other channels such as, e.g., the possibility that workers might be more strongly influenced, *ceteris paribus*, by a reference worker that they have chosen themselves ("*choice effect*").

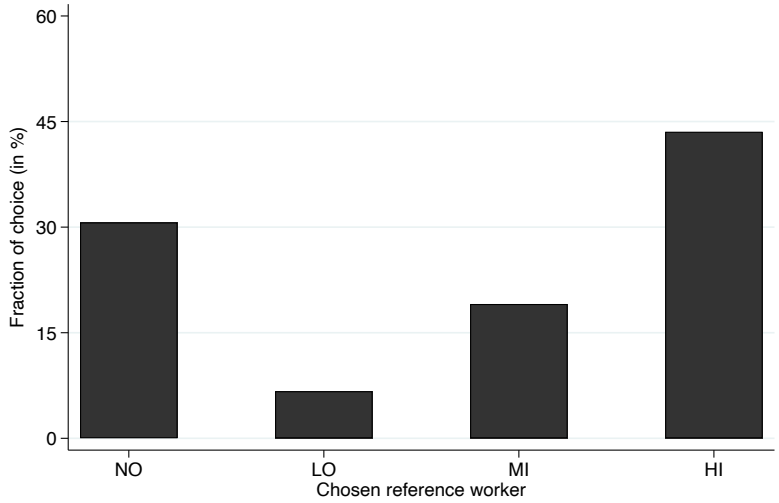
4.3.1 Analyzing preferences for peers and their determinants

Figure 4 depicts the relative frequency with which participants in ENDO choose each of the four available alternatives. The most frequently chosen option is the best performing reference worker (HI; 43 percent), followed by the choice not to compare to any reference worker (NO; 31 percent). The other two alternatives are chosen less frequently (19 percent for MI and 7 percent for LO, 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.[25] A Pearson χ^2 test also rules out that the distribution of chosen reference workers matches the distribution of predicted most

[25]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 relative performance, but slightly underestimate absolute performance of the different reference workers in period 2.

motivating reference workers in EXBE ($p < 0.01$).

Figure 4: 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.

In Appendix E, we shed light on the determinants of participants' choices by exploring how their own productivity in period 1 as well as their gender affects whom they decide to compare to. We document two important findings. First, irrespective of their own productivity in the first round, there is always a substantial proportion of workers who choose not to compare to any reference worker in the second round. Second, among participants who choose to compare to a reference worker, most participants choose a reference worker whose performance is similar to or higher than their own performance: the least productive workers tend to compare to the low productivity reference worker while the most productive ones tend to compare to the high productivity reference workers. 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. Choice patterns are, however, very similar across genders.

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 three independent raters to code participants' answers. Raters were

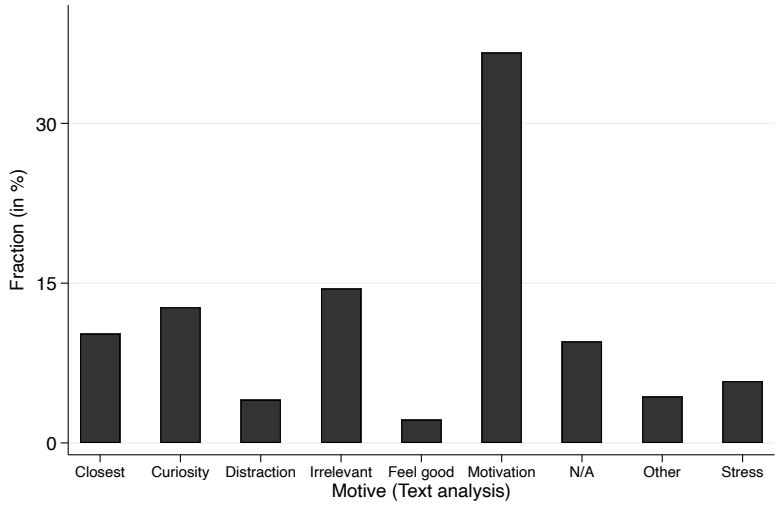
given a list of eight possible choice motives (which were identified through focus groups). Each rater was then asked to assign up to 3 different motives per worker. For example, the answer "*I chose to compare to this reference worker because it was the closest to me and I thought it would motivate 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.[26]

Figure 5 depicts the distribution of choice motives across all participants in ENDO. 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. 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"), 10.25 percent report that the they chose the reference worker that was "Closest to them", 12.73 percent indicate that the choice was made out of "Curiosity", and 5.8 percent directly refer to "Stress" as a key driver of their choice. Last, a minority of 2.17 percent indicate that they chose whatever made them "feel good about themselves."

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 declare 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

[26]We describe the details of the procedure for the text analysis in Appendix C.

Figure 5: 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).

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 E.1 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 differs 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 implies that fewer participants in ENDO

observe the peer that tends to be perceived as most stressful, which is well aligned with the observation that stress increases much less strongly in ENDO than in EXBE. However, it also 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. We provide an answer to this question in the next section.

4.3.2 Exploring the mismatch effect

A key distinction between exogenous assignment and endogenous choice lies in the potential for mismatches: workers who are exogenously assigned a reference worker may be forced to observe someone they would not have chosen themselves. In the EXBE condition, 92% of subjects are matched with the most productive reference worker, but only 41% of these workers identified that peer as their preferred choice—a rate comparable to the ENDO condition (44%). This means that a substantial share of workers in EXBE were paired with a reference worker *that they would not have chosen*.

In this section, we examine the consequences of these mismatches. To that end, we cannot simply compare workers in EXBE who were assigned to the most productive reference worker against their will with those who were assigned to it and wanted to see it. The reason is that workers's preferences for a peer are endogenous and may correlate with other relevant worker characteristics, e.g. productivity. However, we can shed light on this question by using data from EXRA, where workers are randomly assigned to a peer. Specifically, we can partition the EXRA sample according to workers' preferred peer. Within these subsamples, participants have identical preferences, and reference workers are randomly assigned. This approach allows us to cleanly identify how different reference workers affect performance and perceived stress, conditional on preferences for observing a particular reference worker.

Figure 6 displays the changes in performance and perceived stress for workers in EXRA, segmented by their preferred reference worker. For each subsample of workers with identical preferences, the figure shows the causal effects of different (randomly assigned) reference workers on performance (left panels) and stress (right panels). The figure reveals a very clear pattern: assigning participants to the most productive reference worker (HI) is most motivating only for workers who wanted to observe this particular reference worker (Panel (g)). In all other subsamples, being

assigned to the most productive peer is not more motivating than the other options (see Panels (a), (c), and (e)). At the same time, however, Panels (b), (d), (f) and (h) indicate that being assigned the most productive peer always yields the largest increase in stress.

This analysis provides a coherent explanation for why ENDO yields performance gains that are comparable to those in EXBE while at the same time generating much lower stress. In ENDO, only 44% of participants choose to observe the most productive peer, while 31% opt to observe no peer at all and 25% select either the low or the average productivity peer. In EXBE, most participants are forced to observe the high-productivity peer, although the high-productivity peer is *not* the preferred peer for most of these participants. For participants assigned to the highly productive peer against their will, performance does not increase more than it would have if they had been matched with their preferred option, i.e., there is a mismatch effect in performance. However, the matching pattern in EXBE strongly increases perceived stress, as being forced to compare to the most productive peer is the most stressful experience both for workers who are mismatched (Panels (b), (d) and (f)) and those who wanted to compare to this reference worker (Panel (h)). As a result, moving from ENDO to EXBE imposes a cost—in terms of substantially higher perceived stress—without generating a sizeable benefit in terms of output.

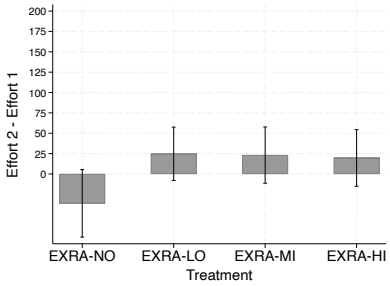
4.3.3 Disentangling and quantifying behavioral mechanisms

What is the role of the composition and the mismatch effect in explaining the differences between EXBE and ENDO? Can their importance be quantified, and can any other behavioral mechanism be identified? In particular, does the mere act of choosing—*ceteris paribus*—have an effect on performance or stress?

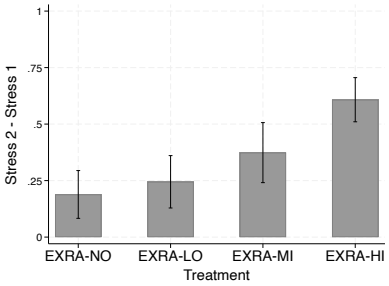
To answer these questions, we conduct a set of Gelbach decompositions (Gelbach, 2016). Conceptually, this approach—which has been widely applied to questions in labor economics (e.g., Cook et al., 2021), political economy (e.g., Stantcheva, 2021), and health economics (e.g., Allcott et al., 2019), among others—treats possible explanatory mechanisms for a given observed (treatment) difference as omitted variables, and allows to assess the size of the bias that would result if each of these mechanisms was excluded. The main advantage of this approach is that it provides

Figure 6: Effects of different exogenously assigned reference worker on productivity (left panels) and stress (right panels), by preferred reference worker (rows)

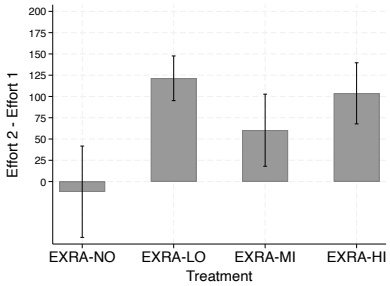
(a) Effort (if Preferred: NO)



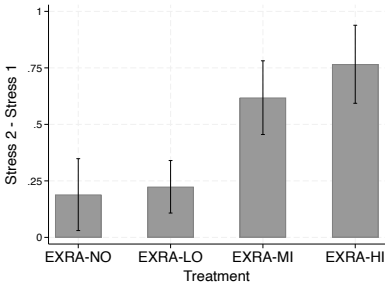
(b) Stress (if Preferred: NO)



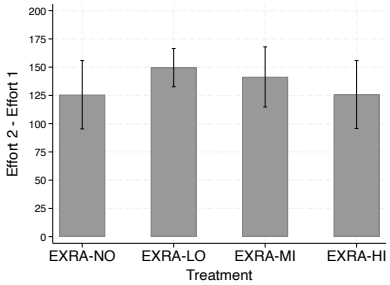
(c) Effort (if Preferred: LO)



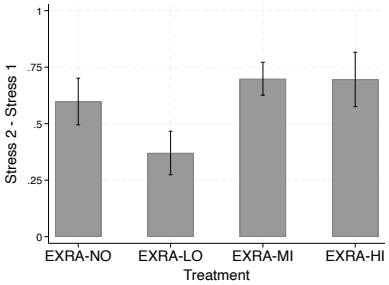
(d) Stress (if Preferred: LO)



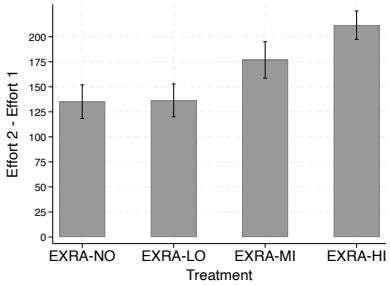
(e) Effort (if Preferred: MI)



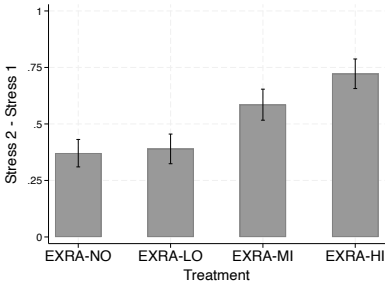
(f) Stress (if Preferred: MI)



(g) Effort (if Preferred: HI)



(h) Stress (if Preferred: HI)



Note: Panels on the left depicts the average change in effort between rounds 1 and 2 in the different EXRA treatments, by preferred reference worker (rows). For example, panel (a) shows the effects of the different randomly assigned reference workers on performance for workers who prefer to compare to no peer (had they had the choice). Panels on the right depict 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 +/- 1 standard error.

an econometrically principled way to identify an exact and interpretable breakdown of the effects of different mechanisms in explaining treatment differences. Moreover, it is invariant to the order in which mechanisms are included[27] and it is well suited to situations in which covariates are intercorrelated (Gelbach, 2016). In our case, we have, for example, documented that preferences for peers correlate with effort in round 1 (see section 4.3.1). As such, it is important to use a method that isolates the effect of preferences for peers without confounding them with the effects of changes in the ability levels of workers. Gelbach decompositions allow to elegantly deal with these empirical challenges.

We are interested in decomposing treatment differences in performance and stress between ENDO and EXBE. In particular, we aim to assess the importance of the composition effect, the mismatch effect, and any remaining variation in explaining these treatment differences. For these purposes, we restrict our attention to the EXBE and the ENDO sample. The Gelbach decomposition compares two models. The first is a so-called "restricted model", where the outcome of interest (e.g., performance) is regressed on a treatment dummy for EXBE, and a set of controls like age and gender. The second is a "full model" that comprises the same set of explanatory variables as the restricted model, but also includes different "mechanism variables". In our application, we include variables aimed at capturing the composition effect and the mismatch effect. To identify the composition effect, we include a set of dummies that indicate which reference worker (if any) participants observe. These dummies allow us to take into account the fact that the matching of participants with reference workers varies across treatments.[28] To capture the mismatch effect, we add a dummy variable that indicates whether participants are observing their preferred reference worker.[29] Together, this set of covariates capture the main behavioral mechanisms

[27] Adding covariates (e.g., mechanisms) sequentially to a regression yields results that are highly dependent on the order in which covariates are added, as demonstrated by Gelbach (2016). It therefore does not allow for an unambiguous and econometrically principled decomposition of treatment differences. For a more detailed discussion of decomposition techniques in economics, see also Fortin et al. (2011).

[28] To capture the fact that the distribution of reference workers varies with effort in round 1 (see Figure E.1 in the Appendix), we also include dummies that interact which reference worker (if any) participants observe with decile dummies for their effort in round 1.

[29] This dummy also takes the value 1 if workers who prefer not to compare with a reference worker are not matched with a reference worker. Note that, by construction, this dummy is always equal to one in ENDO, while it can either take the value 1 or 0 in EXBE.

discussed in the previous sections.

The Gelbach decomposition allows us to assess how much of the change in the treatment dummy between the restricted and the full model is attributable to the different "mechanisms variables". It also quantifies the remaining unexplained variation, which might—depending on the context—be interpretable. Because our Gelbach decomposition controls for both the composition and the mismatch effects, the residual variation can be attributed to a combination of a choice effect[30] and any additional unobserved factor that systematically varies between treatments.

We performed this analysis to explain treatment differences both in performance and in stress. We present the results in Table 2 below, where column (1) depicts the decomposition for performance and column (2) depicts the decomposition for stress. In order to make things more easily comparable across the two dimensions, we express the decomposition in percent of standard deviation of the outcome of interest.

Table 2: Contributions of different mechanisms in explaining differences between ENDO and EXBE.

	Productivity (% of SD)	**Stress (% of SD)**
Composition effect	+20%	+11%
Mismatch effect	-30%	-3%
Residual variation	+13%	+4%
Total effect of EXBE (relative to ENDO)	**+3%**	**+13%**

Note: The table depicts the importance of the different mechanisms in explaining differences in outcome between the ENDO and the EXBE treatments. To make things easily comparable, they are expressed in percent of a standard deviation of the outcome of interest.

We first consider the results for the performance dimension (column 1). The table reveals three striking findings. First, a substantial part of the productivity differences between ENDO and EXBE can be explained by changes in the composition of the reference workers. In fact, the change in the composition of the reference workers contribute to an *increase* in performance in EXBE (relative to ENDO) of about 20% of a standard deviation. This is not surprising as in EXBE most participants are matched

[30] A preference for chosen alternatives, or 'choice effect', has been discussed in other contexts and relates to preferences for autonomy, decision rights and self-determination (see, e.g., Dal Bó et al., 2010; Bartling et al., 2014; Owens et al., 2014; Deci and Ryan, 2013; Bartling et al., 2013; Falk and Kosfeld, 2006).

with the high productivity reference worker, whereas in ENDO a large share of the population chooses *not* to compare with this peer.

Second, we find that the increase in performance in EXBE that is explained by the composition effect is completely offset by a large mismatch effect, which contributes *negatively* to the effect of EXBE (-30% of a standard deviation). Indeed, an advantage of ENDO is that these mismatches are absent since participants always get to observe they preferred peer.

Third, the residual variation is the smallest, positive, contributor to the total effect of EXBE (+13% of a standard deviation). Because the residual variation captures a combination of choice effect and any additional unobserved factor that systematically varies between ENDO and EXBE, these results suggest that the choice effect is the smallest driver of treatment differences and that it is—if anything—negative for performance (i.e., it negatively affects ENDO).

Turning to the decomposition for stress, we find that the change in the composition of the reference worker is the largest contributor to the differences in stress between ENDO and EXBE (+11% of a standard deviation). This result reflects the fact that there are much more workers who compare with the highly productive and highly stressful peer in EXBE. We find virtually no evidence for a mismatch effect in the stress dimension. This is consistent with the evidence presented in Section 4.3.2, where we showed that what matters for stress is whom one observes but not whether one wanted to observe that peer.[31] Last, we find that the residual variation is also negligible, suggesting no choice effect in the stress domain.

Taken together, these results suggest that the absence of a performance difference between EXBE and ENDO stems from a large mismatch effect that fully offsets the composition effect. Although substantially more participants observe the highly productive peer in EXBE, overall performance does not improve, as those assigned to this peer against their preference fail to benefit from a strong motivational spillover. In the domain of stress, by contrast, the composition effect is not neutralized by a mismatch effect. The large number of participants exposed to the most productive peer experience a significant increase in stress, regardless of whether this peer aligns with

[31]Not comparing to a peer or comparing to a low productivity peer is always the least stressful, whereas comparing to the highly productive peer is always the most stressful.

their initial preference. As a result, stress levels rise more in EXBE than in ENDO.

Summarizing, our results reveal that endogenous choice 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.

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 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 address these questions, 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.[32]

We report the effects of these treatments in Appendix B.5. Three important insights emerge from this analysis. First, participants respond to financial incentives as predicted by 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 percent larger than the increase in productivity of the workers in the equivalent treatments without the piece rate (i.e. pooling RANK,

[32]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).

ENDO and EXBE; $p < 0.01$). While these financial incentives have positive effects on performance, they also generate a significant increase in stress of about 13 percent ($p < 0.01$).[33]

Second, 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.4 in Appendix B.5), 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×$) generates an increase in productivity that is of roughly the same magnitude as when forcing them to compare to the most motivating reference worker (EXBE×$)[34] but it also yields a substantially smaller increase in stress ($p < 0.05$). Overall, these results suggest that our findings are not driven by the specificities of a particular compensation scheme, and that social incentives and monetary rewards act as complements in our setting.

[33]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.

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

4.4.2 Non-social comparisons

How important is it for our results that comparisons are 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 goals would generate similar effects.[35] We answer this question by conducting an additional pre-registered experiment, in which we compare participants who observe a highly productive reference worker with participants who are confronted with an equally challenging non-social pacemaker.[36]

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 about the same number of points as the reference worker in the EXRA-HI treatment).[37] 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.

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 similar effects. Alternatively, social comparisons might be more motivating and more stressful than equally challenging non-social goals.

Our results unambiguously show that motivational effects triggered by social comparisons surpass those brought about by otherwise similar 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.13 in Appendix B.7). Social comparisons not only have larger effects on workers' performance, they also

[35] 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 differences reported throughout the paper.

[36] We preregistered this study on AsPredicted.org (trial 137539). For details on the design, see Appendix B.7.

[37] To avoid that participants are suspicious, we rounded up the performance of the pacemaker to 1600 (instead of 1583 for the highly productive peer).

have very different effects on workers' perceptions (see Tables B.13 and B.14 in Appendix B.7). 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

We have shown that social comparisons entail a potentially important tradeoff: observing a peer not only creates motivational spillovers, but it also increases the stress level of the observer. Peer assignment mechanisms importantly shape the behavioral effects of such 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 stress than exogenous assignment to the predicted most motivating peer.

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 has focused on how to best incentivize the workforce, it has often neglected to evaluate the impacts of different incentive schemes for important outcomes such as workers' stress,

or satisfaction. Whether and how companies should trade-off these dimensions is a particularly exciting open question for future research.

Our findings also have useful implications for theory. To reflect the impact of social interactions on effort choices in organizations, Ashraf and Bandiera (2018) propose to extend the standard principal-agent model by incorporating a social interaction term in the agents' utility function. Our empirical evidence provides guidance on how to characterize this element in work environments that share the main characteristics of our task. In Appendix D, we provide a more detailed discussion of Ashraf and Bandiera (2018)'s framework and discuss ways to formally adapt their model on the basis of our findings. We demonstrate that such a stylized model is capable of generating the core patterns observed in our data.

Our experimental design was purposefully kept 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. Moreover, at the workplace, people may not only observe the productivity but also many other characteristics of their peers. Whether and how these elements interact with social comparisons remains an open question which is beyond the scope of our paper. We see our study as an important first step towards addressing these exciting open questions.

Our experimental paradigm could also easily be applied to other contexts where "observing others" is believed to be an important driver of behavior. Previous work has shown that *static* or *aggregate* information about peers can affect behavior in, e.g., public goods (Chen et al., 2010), financial decision making (see, e.g., Kirchler et al., 2018; Schwerter, 2019), labor market decisions (Coffman et al., 2017), and energy consumption (see, e.g., Allcott and Kessler, 2019). Our setup paves the way for tailoring interventions that provide *individualized* and *real-time* information about peers.

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 and mainly limited to educational contexts (see, e.g., Clark and Senik, 2010; Cicala et al., 2018; Kiessling et al., 2019). 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.

_ , Rebecca Diamond, Jean-Pierre Dubé, Jessie Handbury, Ilya Rahkovsky, and Molly Schnell, "Food deserts and the causes of nutritional inequality," *The Quarterly Journal of Economics*, 2019, *134* (4), 1793–1844.

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.

Azmat, Ghazala and Nagore Iriberri, "The provision of relative performance feedback: An analysis of performance and satisfaction," *Journal of Economics & Management Strategy*, 2016, *25* (1), 77–110.

Bandiera, Oriana, Iwan Barankay, and Imran Rasul, "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.

Bartling, Björn, Ernst Fehr, and Holger Herz, "The intrinsic value of decision rights," *Econometrica*, 2014, *82* (6), 2005–2039.

_ , _ , and Klaus M Schmidt, "Discretion, productivity, and work satisfaction," *Journal of Institutional and Theoretical Economics (JITE)/Zeitschrift für die gesamte Staatswissenschaft*, 2013, pp. 4–22.

Bellemare, Charles, Patrick Lepage, and Bruce Shearer, "Peer pressure, incentives, and gender: An experimental analysis of motivation in the workplace," *Labour Economics*, 2010, *17* (1), 276–283.

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.

Bianchi, Renzo, Irvin S Schonfeld, and Eric Laurent, "Physician burnout is better conceptualised as depression," *The Lancet*, 2017, *389* (10077), 1397–1398.

Biasi, Barbara, Michael S Dahl, and Petra Moser, "Career effects of mental health," Technical Report, National Bureau of Economic Research 2021.

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.

Braghieri, Luca, Ro'ee Levy, and Alexey Makarin, "Social media and mental health," *American Economic Review*, 2022, *112* (11), 3660–3693.

Bubonya, Melisa, Deborah A Cobb-Clark, and Mark Wooden, "Mental health and productivity at work: Does what you do matter?," *Labour economics*, 2017, *46*, 150–165.

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.

Bütikofer, Aline, Rita Ginja, Katrine V Løken, and Fanny Landaud, "Higher-Achievement Schools, Peers and Mental Health," *The Economic Journal*, 2023, *133* (655), 2580–2613.

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.

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.

Cobb-Clark, Deborah A, Sarah C Dahmann, and Nathan Kettlewell, "Depression, risk preferences, and risk-taking behavior," *Journal of Human Resources*, 2022, *57* (5), 1566–1604.

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.

Cook, Cody, Rebecca Diamond, Jonathan V Hall, John A List, and Paul Oyer, "The gender earnings gap in the gig economy: Evidence from over a million rideshare drivers," *The Review of Economic Studies*, 2021, *88* (5), 2210–2238.

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.

Currie, Janet and Mark Stabile, "Child mental health and human capital accumulation: the case of ADHD," *Journal of health economics*, 2006, *25* (6), 1094–1118.

Deci, Edward L and Richard M Ryan, *Intrinsic motivation and self-determination in human behavior*, Springer Science & Business Media, 2013.

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

Doerrenberg, Philipp, Denvil Duncan, and Danyang Li, "The (in) visible hand: do workers discriminate against employers?," *Journal of Public Economics*, 2024, *231*, 105065.

Drouvelis, Michalis and Paola Paiardini, "Feedback quality and performance in organisations," *The Leadership Quarterly*, 2022, *33* (6), 101534.

Eisenberg, Daniel, Ezra Golberstein, and Justin B Hunt, "Mental health and academic success in college," *The BE journal of economic analysis & policy*, 2009, *9* (1).

Eriksson, Tor, Anders Poulsen, and Marie Claire Villeval, "Feedback and incentives: Experimental evidence," *Labour Economics*, 2009, *16* (6), 679–688.

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

Ettner, Susan L, Richard G Frank, and Ronald C Kessler, "The impact of psychiatric disorders on labor market outcomes," *ILR Review*, 1997, *51* (1), 64–81.

Falk, A. and M. Kosfeld, "The hidden costs of control," *The American economic review*, 2006, pp. 1611–1630.

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.

Fletcher, Jason M, "Adolescent depression and educational attainment: results using sibling fixed effects," *Health economics*, 2010, *19* (7), 855–871.

_ , "The effects of childhood ADHD on adult labor market outcomes," *Health economics*, 2014, *23* (2), 159–181.

Fortin, Nicole, Thomas Lemieux, and Sergio Firpo, "Decomposition methods in economics," in "Handbook of labor economics," Vol. 4, Elsevier, 2011, pp. 1–102.

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.

Gelbach, Jonah B, "When do covariates matter? And which ones, and how much?," *Journal of Labor Economics*, 2016, *34* (2), 509–543.

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.

Hayes, Andrew F and Klaus Krippendorff, "Answering the call for a standard reliability measure for coding data," *Communication methods and measures*, 2007, *1* (1), 77–89.

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.

Kessler, Judd and Lise Vesterlund, "The external validity of laboratory experiments: The misleading emphasis on quantitative effects," *Handbook of experimental economic methodology*, 2015, *18*, 392–405.

Kiessling, Lukas and Jonathan Norris, "The long-run effects of peers on mental health," *The Economic Journal*, 2023, *133* (649), 281–322.

_ , 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.

Krippendorff, Klaus, "Measuring the reliability of qualitative text analysis data," *Quality and quantity*, 2004, *38*, 787–800.

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.

Kuhnen, Camelia M and Agnieszka Tymula, "Feedback, self-esteem, and performance in organizations," *Management Science*, 2012, *58* (1), 94–113.

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.

Owens, David, Zachary Grossman, and Ryan Fackler, "The control premium: A preference for payoff autonomy," *American Economic Journal: Microeconomics*, 2014, *6* (4), 138–161.

Quidt, Jonathan De and Johannes Haushofer, "Depression for economists," Technical Report, National Bureau of Economic Research 2016.

_ , _ , and Christopher Roth, "Measuring and bounding experimenter demand," *American Economic Review*, 2018, *108* (11), 3266–3302.

Reiff, Joseph S, Justin C Zhang, Jana Gallus, Hengchen Dai, Nathaniel M Pedley, Sitaram Vangala, Richard K Leuchter, Gregory Goshgarian, Craig R Fox, Maria Han et al., "When peer comparison information harms physician well-being," *Proceedings of the National Academy of Sciences*, 2022, *119* (29), e2121730119.

Ridley, Matthew, Gautam Rao, Frank Schilbach, and Vikram Patel, "Poverty, depression, and anxiety: Causal evidence and mechanisms," *Science*, 2020, *370* (6522), eaay0214.

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

Rosaz, Julie, Robert Slonim, and Marie Claire Villeval, "Quitting and peer effects at work," *Labour Economics*, 2016, *39*, 55–67.

Roth, Christopher, Peter Schwardmann, and Egon Tripodi, "Depression stigma," 2024.

_ , _ , and _ , "Misperceived effectiveness and the demand for psychotherapy," *Journal of Public Economics*, 2024.

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

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

Stantcheva, Stefanie, "Understanding tax policy: How do people reason?," *The Quarterly Journal of Economics*, 2021, *136* (4), 2309–2369.

Stewart, Walter F, Judith A Ricci, Elsbeth Chee, Steven R Hahn, and David Morganstein, "Cost of lost productive work time among US workers with depression," *Jama*, 2003, *289* (23), 3135–3144.

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.

ONLINE APPENDIX

A Details on the experiment and the population sample

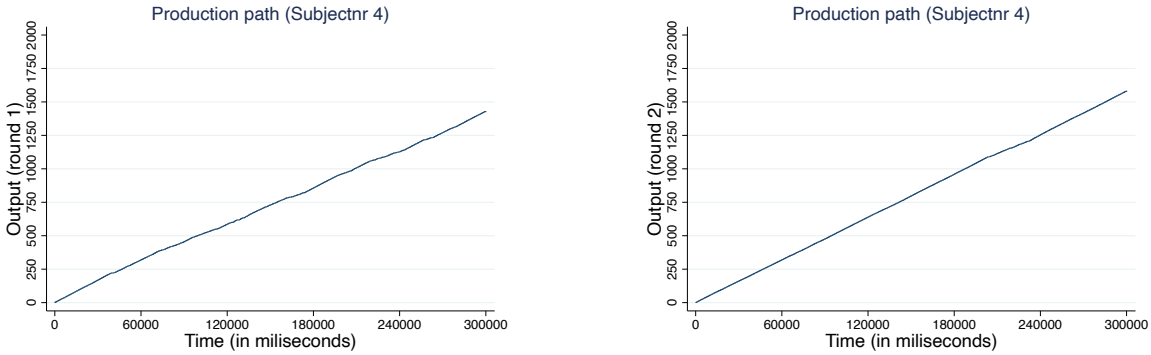
A.1 Reference population and potential reference workers

Table A.1 depicts the ranking (rank) of the 60 workers from the reference population, sorted by their output in round 1 (effort1). The three potential reference workers (rank 4, 26 and 49) are highlighted **bold**. We selected these three workers based on the following criteria: a) they 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. Figures A.1 to A.3 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**)

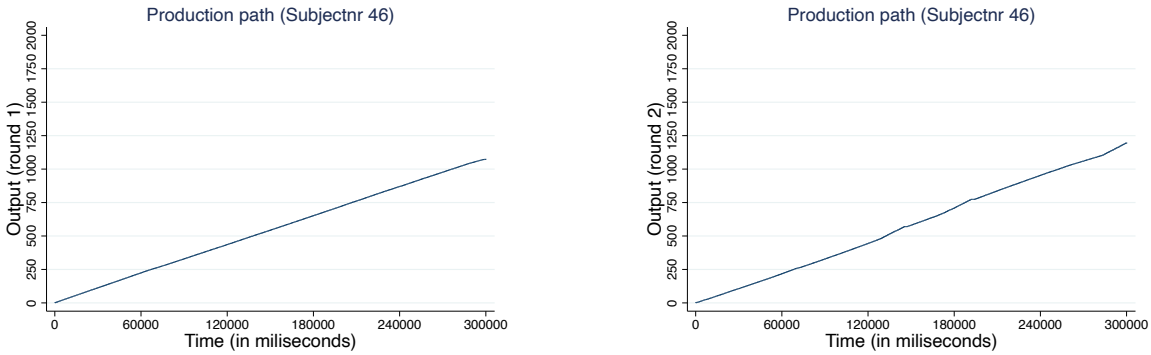
rank	subjectnr	effort1	effort2
1	36	1553	1128
2	25	1488	1474
3	6	1446	1458
4	**4**	**1428**	**1580**
5	13	1415	1048
6	39	1409	1426
7	15	1366	544
8	19	1338	519
9	27	1325	1231
10	16	1307	1338
11	42	1301	1016
12	18	1299	1300
13	55	1284	1244
14	23	1259	861
15	20	1249	1226
16	3	1238	1081
17	51	1231	1326
18	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.1: Production paths for the high productivity reference worker (HI, subjectnr=4, rank=4)



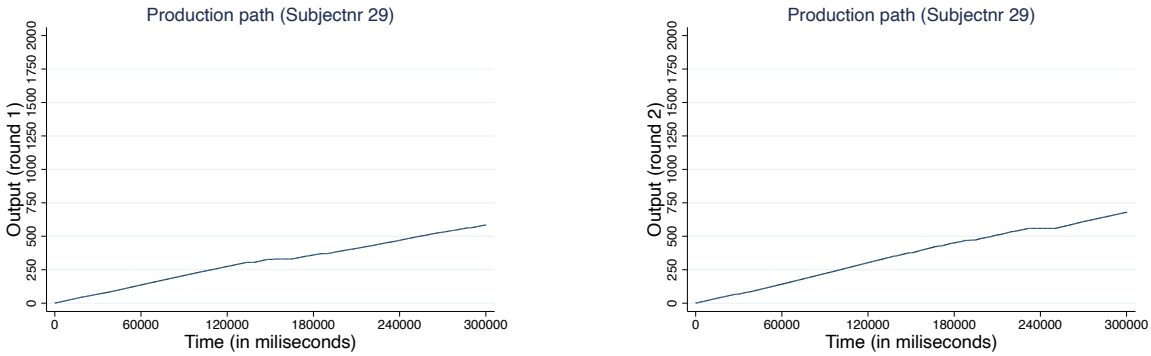
(a) Round 1 (b) Round 2

Figure A.2: Production paths for the average productivity reference worker (MI, subjectnr=46, rank=26)



(a) Round 1 (b) Round 2

Figure A.3: Production paths for the low productivity worker (LO, subjectnr=29, rank=49)



(a) Round 1 (b) Round 2

A.2 Description of the different treatments and sample size

In Figure A.4, we provide an overview of all the treatments implemented in the main experiment. In Table A.2, we depict the key characteristics of each treatment (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. Details on the sample for the follow-up experiment are provided in Appendix B.7.

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.

Figure A.4: Overview of the full 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)

A.3 Mturk, experimental protocol, and eligibility criteria

We ran our experiment on Amazon Mechanical Turk (MTurk). MTurk is an online labor market where employers can advertise small jobs ('HITs') that typically consist of simple, repetitive tasks.[38] 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 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). 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 for worker motivation.

We required that workers are US residents, that they have an approval rate of at least 95%, and a minimum of 50 approved tasks. 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. everything that workers see during the first half of the study (including the HIT description) is the same across all treatments. This prevents workers with different characteristics selecting 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[39], (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.[40] The final sample includes 6532 subjects.[41]

[38] 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.

[39] 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.

[40] 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.

[41] 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.2.

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 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 RANK. For all 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 Material related to the estimation of treatment effects

This Appendix contains the material related to the estimation of the treatment effects. 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 (e.g., 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 two treatments, 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 residuals 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 reveals 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. The effects of the different treatments for stress are estimated with a similar procedure.

[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 significant 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:

$$\begin{aligned}\text{effort}_{it} =& \beta_1 \text{EXRA-HI}_i + \beta_2 \text{EXRA-MI}_i + \beta_3 \text{EXRA-LO}_i + \beta_4 \text{EXRA-NO}_i + \beta_5 \text{RANK}_i \\ &+ \beta_6(\text{EXRA-HI}_i \times \text{P2}_t) + \beta_7(\text{EXRA-MI}_i \times \text{P2}_t) + ... + \beta_{11}(\text{RANK}_i \times \text{P2}_t) + \epsilon_{it}\end{aligned}$$

We report the results in the Table B.1 below. "EXRA-HI x P2" shows the motivational effect of being assigned to the EXRA-HI treatment (β_6), i.e. by how much production increases from period 1 to period 2 in the EXRA-HI treatment. Similarly, "RANK x P2" shows the motivational effect of the RANK treatment (β_{11}), i.e. by how much production increases from period 1 to period 2 in the RANK treatment. The baseline productivity levels (β_1 to β_5 are very similar across treatments by virtue of randomization and are therefore not reported, see "Treatment dummies"). This table shows that the increase in performance is the largest in EXRA-HI (+ 148.04 units of output, $p < 0.01$). The different test of equality of coefficients are reported at the bottom of the table.

We use a similar procedure to assess the effects of different treatments for stress. All the effects discussed throughout the paper are assessed in a similar way.

Table B.1: 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
Ho: RANK x P2 = EXRA-NO x P2	0.330	0.330	0.036	0.036
Ho: RANK x P2 = EXRA-LO x P2	0.003	0.003	0.009	0.009
Ho: RANK x P2 = EXRA-MI x P2	0.000	0.000	0.064	0.064
Ho: RANK x P2 = EXRA-HI x P2	0.000	0.000	0.000	0.000
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.2.2 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 ($e_1(MI) \leq e_1 \leq e_1(HI)$)

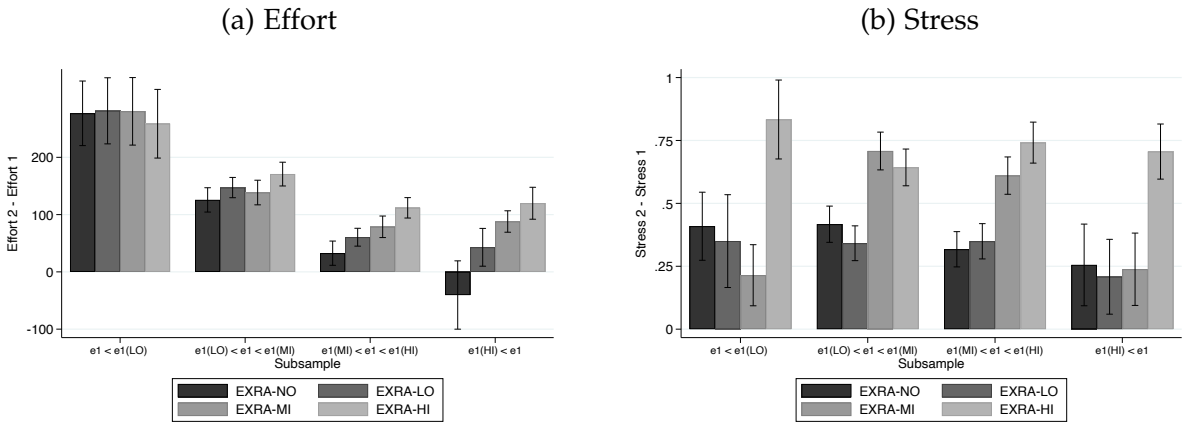
d) workers with an output in period 1 that is *higher* than the output in period 1 of the high productivity worker ($e_1(HI) \leq e_1$)

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.2 and B.3). 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 111.84 units when exogenously assigned to HI ($p < 0.01$), by 78.70 when assigned to MI ($p < 0.1$) and by 60.53 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.3).

In Tables B.4 to B.7, 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.

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). Whiskers represent +/- 1 standard error.

Table B.2: 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.3: 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

Table B.4: 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.5: 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.6: 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.7: 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.8: 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 Stress-adjusted output

Figure B.2 depicts stress-adjusted output, defined as effort in period 2 divided by stress in period 2, as a function of treatments, and Table B.9 depicts the associated regressions. This analysis confirms that stress-adjusted output is significantly higher in ENDO than in EXBE ($p < 0.01$), i.e., that output increases faster than stress in ENDO. These results are robust to defining stress-adjusted output as effort2/stress2 – effort1/stress1 (See Figure B.3 and Table B.10). There too, stress-adjusted output is higher in ENDO than in EXBE (although the statistical significance of the result is weaker, $p = 0.08$).

Figure B.2: Stress-adjusted output (by treatment)

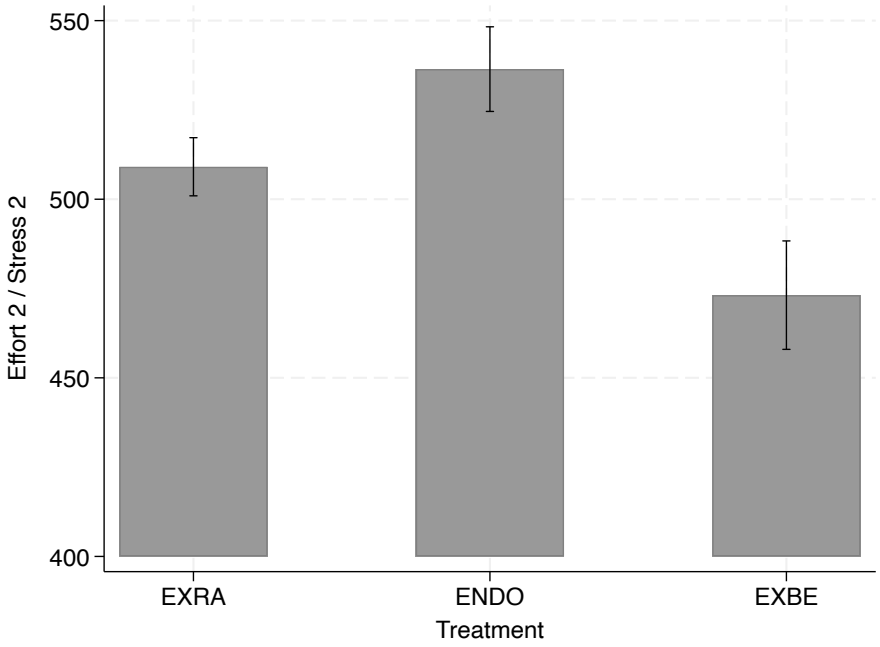


Table B.9

	Stress-adjusted output (1)	(2)
ENDO	27.338*	30.185**
	(14.375)	(14.410)
EXBE	-35.942**	-33.721**
	(17.217)	(17.158)
Male (=1)		44.845***
		(12.560)
Age		-1.468***
		(0.464)
Constant	509.090***	541.073***
	(8.141)	(20.374)
Ho: ENDO=EXBE	0.001	0.001
R^2	0.003	0.010
Observations	3530	3530

Note: OLS estimations. Dependent variable is stress-adjusted output (efffort2/stress2). Baseline category are subjects in the EXRA treatment. Levels of significance: $*p < 0.1, **p < 0.05, ***p < 0.01$.

Figure B.3

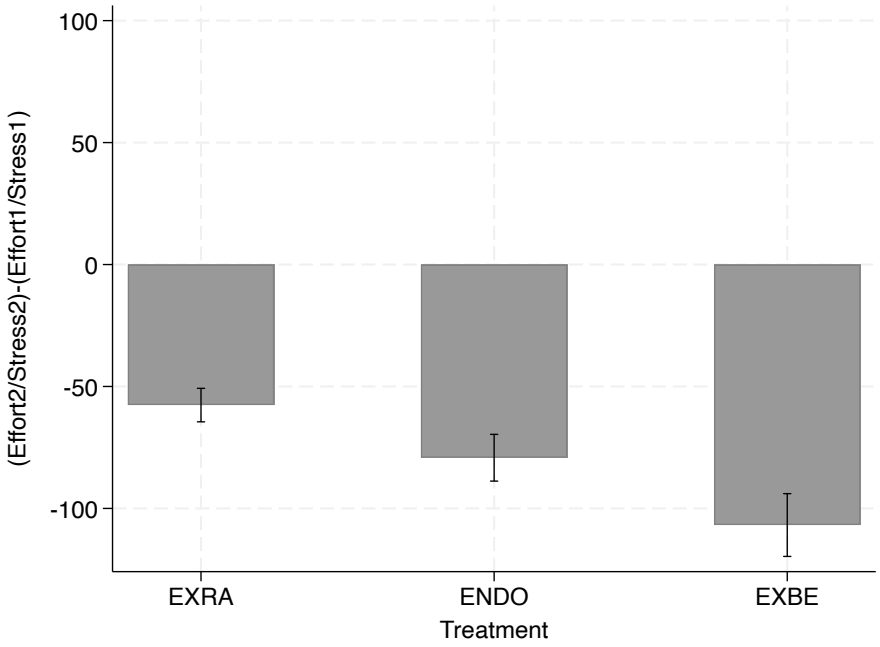


Table B.10: Regressions

	(Effort2/Stress2) - (Effort1/Stress1)	
	(1)	(2)
ENDO	-21.590*	-20.526*
	(11.789)	(11.801)
EXBE	-49.183***	-47.819***
	(14.582)	(14.523)
Male (=1)		11.651
		(10.415)
Age		1.498***
		(0.365)
Constant	-57.626***	-117.886***
	(6.862)	(16.669)
ENDO=EXBE	0.086	0.088
R^2	0.003	0.007
Observations	3529	3529

OLS estimations. Baseline category are subjects in the EXRA treatment. Levels of significance: $*p < 0.1, **p < 0.05, ***p < 0.01$.

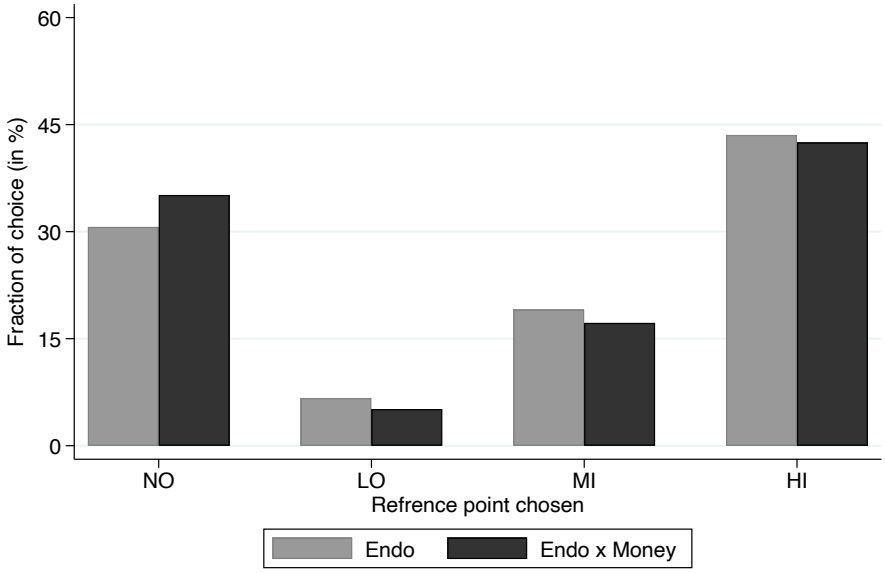
B.5 Benchmarking and robustness

Table B.11: 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 interaction 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: $^{*}p < 0.1, ^{**}p < 0.05, ^{***}p < 0.01$.

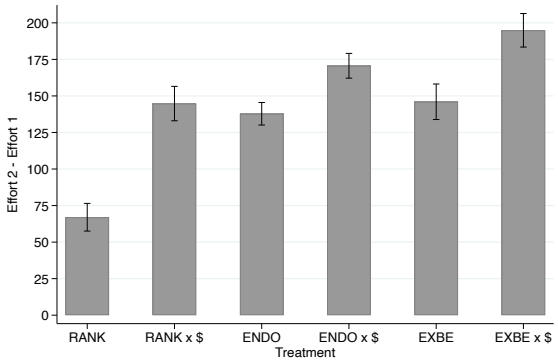
Figure B.4: The effects of monetary incentives on the distribution of chosen reference workers



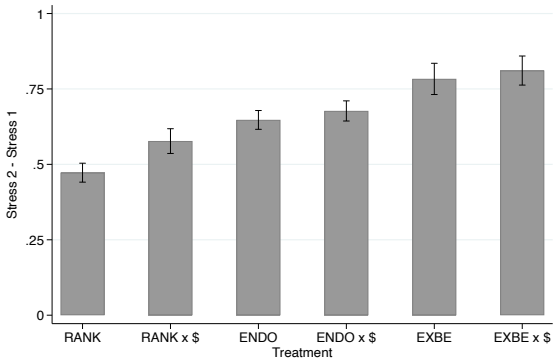
Note: Grey bars represent the distribution of choices for the different reference workers in ENDO. Black bars represent the distribution of choices in ENDO×$.

Figure B.5: The effects of monetary incentives and social comparisons

(a) Effort



(b) Stress



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 +/- 1 standard error.

B.6 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.2.2 (Tables B.4 and B.6) as a basis for our predictions.

Our rule for this tailored exogenous matching is therefore:

- If the participant has an output 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 output 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.2.2 did *not* reveal any gender differences in participants' responses to the different exogenously assigned reference workers.

B.7 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] The instructions are available in Appendix G.5.

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 about the same number of points as the reference worker in the EXRA-HI treatment.[46] 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.

In Table B.12, 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.13 show that the motivational spillovers are much larger in the EXRA-HI condition (+137 points, $p < 0.01$) than in the PACE-HI condition (+64 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.13) and more nervous ($p < 0.01$, column 1 of Table B.14) 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.14, $p < 0.01$), ii) generated a greater feeling of competition (column 3 of Table B.14, $p < 0.01$), and iii) positively affected their performance (column 4 of Table B.14, $p < 0.01$).

[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. Likewise, we assign a few subjects to a randomly assigned reference worker that it *not* the most productive. These observations are irrelevant for our analysis and we therefore don't discuss them here (as we pre-registered).

[46] To avoid that participants are suspicious, we rounded up the performance of the fast pacemaker to 1600 (instead of 1583 for the highly productive peer).

Table B.12: 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.13: 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.14: 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 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.[47] 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."

It turns out that raters almost never assigned three motives to a particular answer. Rater #1 *never* used this option, and Raters #2 and #3 used it only 3 and 12 times (out of 1916 open-ended answers to rate), respectively. For this reason, we focus our analysis on the first two motives identified by the raters. With this procedure, we obtain a maximum of six possible motives (3 raters × 2 possible motives) per rated answer. We aggregate these assessments at

[47] 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.*"

the worker-level by extracting the modal motive, i.e. the motive that is most often identified across raters. This procedure aligns with the recommendations of Krippendorff (2004) to use at least three coders and to rely on a majority decision as a "formal decision rule" to assign final codes. 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.[48]

Overall, there is a generally high degree of agreement between the raters. Across all the responses that had to be rated, there was an agreement on the motive between at least two raters 91.02% of the time, i.e., more than nine times out of ten, at least two raters pointed out the same motive. As a matter of fact, there is perfect agreement between all the raters in 73.33% of the cases, where the same motive is identified by all the raters.[49]

[48]There is no unique mode in about 8% of the cases. This can happen if, for example, all the raters bring up different motives, or if different motives are brought up equally often (e.g., if two motives are brought up three times, or if three motives are brought up two times).

[49]In addition, we also computed Krippendorff's alpha, a widely used measure of intercoder-agreement which is considered to be the most conservative reliability measure (Hayes and Krippendorff, 2007). This analysis shows that intercoder agreement is generally very high, in particular for all the frequenty assigned motives.

D Implications for Theory

Our findings provide useful insights for theory. In this Appendix, we analytically illustrate how a principal-agent model that incorporates utility from social comparisons can generate predictions that are qualitatively aligned with the main empirical regularities documented in our study. To that end, we build on Ashraf and Bandiera (2018) who review the evidence on the impact of social interactions on effort choices in organizations. They suggest that the standard principal-agent model needs adjustment to account for these social interactions and they propose adding a social interaction term to the agent's utility function. In their discussion, they point out that this social interaction term can take various forms (depending on the nature of social interactions and the specific characteristics of the work environment). In section D.1, we describe a social interaction term that captures the key elements of the social spillovers that we observe in our setting. In section D.2, we discuss the implications of such a model for effort provision and stress. Despite the highly simplified and stylized nature of our framework, we show that it is able to generate qualitative predictions that are aligned with our main empirical findings.

D.1 Model Setup

D.1.1 Worker's Utility Function

The general utility function suggested by Ashraf and Bandiera (2018) takes the following form:

$$U_i = m(y_i) - c(y_i) + s(y_i, y_j, \hat{y}_j),$$

where y_i is the worker's own performance, y_j is the performance of the observed peer, $\hat{y}_j$ is the performance of the *preferred* peer[50], $m(\cdot)$ captures all material benefits and direct intrinsic utility that the worker obtains from performing the task, $c(\cdot)$ is the worker's cost of performance (effort), and $s(\cdot)$ is the utility the worker derives from social comparisons.[51]

Our data indicate that in the type of work environment we examine, social comparisons may generate motivational spillovers and affect perceived stress. The social interaction term $s(\cdot)$ should therefore capture both elements. Inter-individual heterogeneity in social perceptions requires that the weights of the different components of social utility are idiosyncratic. To capture this aspect, we introduce an individual-specific weight for the stress component of the social comparison term:

[50] We use the notation $y_j = 0$ for the cases in which the worker does not observe a peer, and $\hat{y}_j = 0$ for the cases in which the worker prefers not to observe a peer.

[51] Generally, $m(\cdot)$ might also depend on y_j (for example, if compensation includes relative performance pay or team incentives). In our study, however, material benefits are independent of peer performance.

$$s(y_i, y_j, \hat{y}_j) = \beta(y_i, y_j, \hat{y}_j) - \delta_i \sigma(y_i, y_j, \hat{y}_j),$$

where $\beta(\cdot)$ stands for the motivational benefits that the worker get from observing a peer, $\sigma(\cdot)$ captures the perceived stress, and δ_i is the individual-specific, relative weight that perceived stress receives in the social interaction term. We consider a worker population characterized by a distribution of preferences $F(\delta)$.

D.1.2 Characteristics of the Components of Social Utility

Now that we defined the general structure of the social interaction term $s(\cdot)$, we can turn to specifying its main properties. We draw on our empirical findings to guide this characterization. In particular, we clarify how the motivational benefits β, and the perceived stress σ, vary as a function of the performance of the observed peer.

Motivational benefits. Our data suggest the following three properties:

1. The marginal social benefit of performance when observing a peer is positive: $\frac{\partial \beta(y_i, y_j \neq 0, \hat{y}_j)}{\partial y_i} > 0$.
2. The marginal social benefit increases in the performance level of the observed peer: $\frac{\partial \beta(y_i, y_j \neq 0, \hat{y}_j)}{\partial y_i \partial y_j} > 0$.
3. The impact of the peer's performance on the worker's marginal social benefit is smaller if the observed peer is not the worker's preferred peer: $\frac{\partial \beta(y_i, y_j = \hat{y}_j)}{\partial y_i \partial y_j} > \frac{\partial \beta(y_i, y_j \neq \hat{y}_j)}{\partial y_i \partial y_j}$.

The first two assumptions suggests that the observer's optimal effort level increases with the observed peer's performance (motivational spillover), while the third assumption reflects the mismatch effect.

Perceived stress. Turning to the effects of social comparisons for *perceived stress*, our data reveal three important insights:

1. Perceived stress depends on the worker's own performance and the performance of the observed peer. It is, however, *not* substantially influenced by the worker's preference for a peer: $\sigma(y_i, y_j, \hat{y}_j) = \sigma(y_i, y_j)$.
2. Perceived stress increases in the performance level of the observed peer: $\frac{\partial \sigma(y_i, y_j)}{\partial y_j} > 0$.
3. Perceived stress is positively correlated with the worker's own performance: $\frac{\partial \sigma(y_i, y_j)}{\partial y_i} > 0$.

The first assumption implies the absence of a mismatch effect in the stress domain. The second and third assumptions relate to the fact that the observer's stress level increases in the performance level of the observed peer.

D.1.3 Solution Concept

We are interested in modeling the aggregate behavior and outcomes for a worker population for the set of peer-assignment mechanisms that we implemented in our study: EXRA, ENDO and EXBE.

In EXRA and EXBE, there are two endogenous variables that each worker chooses: her preferred peer ($\hat{y}_j$)—this choice includes the preference not to observe any peer—and her individual performance (y_i). In these two treatments, whether or not the worker gets to observe her preferred peer depends on the realization of the exogenous peer assignment. In ENDO, the worker picks her preferred peer (if any), which she can observe with certainty ($y_j = \hat{y}_j$).

We assume that workers go through the following choice sequence when determining their optimal behavior:

1. Workers first determine their preferred peer. This choice explicitly includes the option not to observe any peer. The preferred peer is defined as the peer that would yield the highest utility of all available peers if the worker had the option to choose. The choice of the preferred peer involves two steps. First, workers anticipate their own optimal performance level for each possible choice of a peer ($y_i^*(y_j = \hat{y}_j) \forall j$). Second, using these optimal performance levels, they evaluate the resulting utilities for each possible peer (including no peer) and then determine the preferred peer as the one that yields the highest utility: $\max_{\hat{y}_j} U_i(y_i^*(y_j = \hat{y}_j), y_j = \hat{y}_j, \hat{y}_j)$.

2. Which peer-assignment mechanism is used to match workers with a peer (y_j) depends on the treatment. In EXRA and EXBE, this process is exogenous and cannot be influenced by the worker. Hence, some workers will end up observing their preferred peer ($y_j = \hat{y}_j$) while others will end up observing a particular peer against their will (i.e., $y_j \neq \hat{y}_j$). In ENDO, the worker chooses the preferred peer determined in the previous step ($y_j = \hat{y}_j$).

3. The worker chooses the optimal performance level given the preferred peer determined in step 1 ($\hat{y}_j$) and the peer that is assigned in step 2 (y_j).

Importantly, this sequence implies that workers cannot strategically adjust their preferred peer ex post.[52] In other words, workers who are *not* matched with their preferred peer cannot increase their utility by tricking themselves into believing that the assigned peer is the one they would have chosen in the first place.

[52] I.e., in EXRA and EXBE, workers cannot decide to switch their preference to the peer that has been exogenously assigned to them if this peer was not the one they initially preferred.

D.2 Simple Example

In the previous section, we outlined the core structure of a stylized principal-agent model enriched with utility considerations stemming from social comparisons. The key modeling assumption on the characteristics of the social component of the utility function, $s(\cdot)$, are based on our empirical findings. In this section, we add functional forms to illustrate that such a model is able to generate qualitative predictions that are in line with the central patterns observed in our data.

D.2.1 Assumptions and Functional Forms

For simplicity, we make the following additional assumptions:

1. There is only one reference worker available. This reference worker's performance level is defined as $y_j = J$. The worker can only choose between observing no peer at all ($\hat{y}_j = 0$) or observing this specific peer ($\hat{y}_j = J$).

2. We assume that all utility components except the cost function are linear functions. The cost function is convex, but has a linear marginal cost function. Specifically, we impose the following functional forms for the different components:

   - $m(y_i) = w + \eta y_i$, where w is a fixed wage and η denotes constant marginal intrinsic utility.
   - $c(y_i) = \frac{1}{2} y_i^2$.
   - $\beta(y_i, y_j, \hat{y}_j) = \begin{cases} b y_j y_i, & \text{if } y_j = \hat{y}_j, \text{where } b > 0, \\ \epsilon y_j y_i, & \text{if } y_j \neq \hat{y}_j, \text{where } \epsilon \to 0^+. \end{cases}$
   - $\sigma(y_i, y_j) = d y_j + f y_i$, where $f > 0$ and $d > 0$.

D.2.2 Determination of the Preferred Peer

To determine its preferred option, the worker needs to compare the utility from desiring to observe the peer and being matched to the peer ($U_i(y_i^*(y_j = \hat{y}_j = J), y_j = J, \hat{y}_j = J)$) to the utility from not desiring to observe the peer and not being matched to the peer ($U_i(y_i^*(y_j = \hat{y}_j = 0), y_j = 0, \hat{y}_j = 0)$).

Not observing a peer. We first consider the simpler case in which the worker desires not to observe a reference worker ($y_j = \hat{y}_j = 0$). In this case, the worker's utility function is:

$$U_i(y_i, y_j = 0, \hat{y}_j = 0) = w + \eta y_i - \frac{1}{2} y_i^2 - \delta_i f y_i.$$

Optimal performance is:

$$y_i^*(y_j = \hat{y}_j = 0) = \eta - \delta_i f. \tag{1}$$

Utility given optimal performance is:

$$U_i(y_i^*(y_j = \hat{y}_j = 0), y_j = 0, \hat{y}_j = 0) = w + \frac{1}{2}(\eta - \delta_i f)^2. \tag{2}$$

Observing a peer. Next we analyze the case in which the worker desires to observe the peer ($\hat{y}_j = J$). In this case, the utility function is:

$$U_i(y_i, y_j = J, \hat{y}_j = J) = w + \eta y_i - \frac{1}{2}y_i^2 + bJy_i - \delta_i(dJ + fy_i).$$

Optimal performance is:

$$y_i^*(y_j = \hat{y}_j = J) = \eta + bJ - \delta_i f. \tag{3}$$

Utility given optimal performance is:

$$U_i(y_i^*(y_j = \hat{y}_j = J), y_j = J, \hat{y}_j = J) = w + \frac{1}{2}(\eta + bJ - \delta_i f)^2 - \delta_i dJ. \tag{4}$$

The worker therefore prefers to observe the peer if:

$$\delta_i < \frac{2\eta b + b^2 J}{2(bf + d)} = \tilde{\delta}.$$

The expression above implies that the worker chooses to observe the peer if the subjective weight on perceived stress for the worker's utility is sufficiently low. This condition is more likely to be satisfied if:

- The worker's marginal intrinsic utility (η) is high, so that an increase in own performance strongly increases the worker's utility.
- The motivational spillover of observing the desired peer (b) is high, so that observing the worker has a strong impact on the worker's performance.
- The productivity of the peer J is high, because more productive peers are more motivating.
- The link between perceived stress and performance (f) is low, because this link reduces the benefit of higher performance.
- The link between the observed peer's performance and perceived stress (d) is weak, because this link makes observing a peer costly in terms of utility.

D.2.3 Performance and Utility with Mismatched Peer Assignment

Whereas workers in ENDO always choose their preferred option, workers in EXRA and EXBE might be exogenously assigned in ways that do not correspond with their preferences. It is

therefore also necessary to analyze how a worker's optimal performance and the resulting utility are determined for those cases.

Mismatch case 1: Observing a peer against one's will. We begin with the case in which the worker desires not to observe a reference worker, but is exogenously matched with a peer ($y_j = J, \hat{y}_j = 0$). In this case the worker's utility function is:

$$U_i(y_i, y_j = J, \hat{y}_j = 0) = w + \eta y_i - \frac{1}{2}y_i^2 + \epsilon J y_i - \delta_i(dJ + f y_i).$$

Optimal performance is:

$$y_i^*(y_j = J, \hat{y}_j = 0) = \eta + \epsilon J - \delta_i f. \tag{5}$$

Utility given optimal performance is:

$$U_i(y_i^*(y_j = J, \hat{y}_j = 0), y_j = J, \hat{y}_j = 0) = w + \frac{1}{2}(\eta + \epsilon J - \delta_i f)^2 - \delta_i dJ. \tag{6}$$

Comparing equation (5) with equation (1) reveals that assigning a peer to a worker who would have preferred *not* to observe the peer has almost no motivating effect, i.e., $y_i^*(y_j = J, \hat{y}_j = 0) \simeq y_i^*(y_j = \hat{y}_j = 0)$.[53] At the same time, however, the worker experiences an increase in stress that is due to being forced to observe the peer against his will.[54]

Mismatch case 2: Observing no peer against one's will. The reversed form of mismatch consists in not assigning the peer to a worker who would have liked to observe a reference worker ($y_j = 0, \hat{y}_j = J$). In this case the utility function is:

$$U_i(y_i, y_j = 0, \hat{y}_j = J) = w + \eta y_i - \frac{1}{2}y_i^2 - \delta_i f y_i.$$

Optimal performance is:

$$y_i^*(y_j = 0, \hat{y}_j = J) = \eta - \delta_i f. \tag{7}$$

Utility given optimal performance is:

$$U_i(y_i^*(y_j = 0, \hat{y}_j = J), y_j = 0, \hat{y}_j = J) = w + \frac{1}{2}(\eta - \delta_i f)^2. \tag{8}$$

Comparing equation (7) with equation (3) reveals that not assigning a peer to a worker who would have preferred to observe a peer has a demotivating effect, i.e., $y_i^*(y_j = 0, \hat{y}_j = J) < y_i^*(y_j = \hat{y}_j = J)$. While the worker benefits from a lower level of stress, this gain does not offset the utility loss from reduced performance. This follows from the fact that we are considering

[53] Recall that $\epsilon \to 0^+$.

[54] This is visible by comparing equation (2) and (6). The increase in stress is reflected by the negative term $-\delta_i dJ$ in equation (6).

a worker who prefers to observe a peer, which in turn requires that the weight placed on stress is sufficiently low: $\delta_i < \tilde{\delta}$.

D.2.4 Predicted Outcomes for Different Peer-Assignment Mechanisms

We can now evaluate the predicted outcomes for the RANK condition and for the different peer-assignment mechanisms implemented in our study: random assignment (EXRA), endogenous choice (ENDO), and targeted matching (EXBE). As we will show, this highly stylized framework makes predictions that are qualitatively aligned with our empirical findings.

To streamline notation, we define $\alpha = F(\tilde{\delta})$ as the share of the worker population who desires to observe the peer. For simplicity, we set $\epsilon = 0$ in what follows. This assumption simply implies that forcing workers to observe a peer against their will has no motivating effect at all.

RANK. In the RANK treatment, workers do not compare to any peers while working on the task. Expected performance is therefore completely determined by their marginal intrinsic motivation (η) and their marginal anticipated stress ($\delta_i f$):

$$E[y] = \eta - E[\delta]f.$$

Since workers do not observe peers, the expected stress level is simply proportional to expected performance:

$$E[\sigma] = [\eta - E[\delta]f]\, f.$$

EXRA: Random Assignment. In this stylized example, random assignment implies that half of the participants in EXRA are randomly chosen to observe the peer. As a consequence, half of the workers with a preference for observing the peer can do so, while the other half is denied this possibility. Likewise, half of the workers who prefer not to observe a peer are not assigned a peer, whereas the other half are forced to observe the peer against their will.

In terms of performance this implies that only the workers who desire to observe the peer and are actually matched with the peer will be be strongly motivated (Equation 3). This group corresponds to a population share of $\frac{\alpha}{2}$. All other workers will "only" perform at the level at which they would also have performed in the absence of the possibility for social comparisons (see RANK). Expected performance can therefore be expressed as follows:

$$E[y] = \eta - E[\delta]f + \frac{\alpha}{2}bJ.$$

The expected perceived stress level of workers is a function of the performance of the poten-

tially observed reference worker and their own performance:

$$E[\sigma] = \frac{dJ}{2} + \left[\eta - E[\delta]f + \frac{\alpha}{2}bJ\right]f.$$

The first term captures the stress created by observing the peer (which is relevant for half the worker population). The second term is proportional to the expected performance.

These predictions are perfectly in line with the behavioral patterns observed in our experiment: when compared to RANK, both expected performance and stress are higher in EXRA.

ENDO: Endogenous Choice. Endogenous choice implies that only people who desire to observe the peer will choose to do so in ENDO (this is true for a population share α). Workers who prefer not to observe a peer will abstain $(1-\alpha)$. As a consequence the matching pattern in ENDO will be very different from the one in EXRA, where only half of the workers who desired to see the peer were "allowed" to do so $(\frac{\alpha}{2})$ and half of the those who would have preferred not to see the peer were nevertheless "forced" to observe.

Since $\alpha > \frac{\alpha}{2}$, endogenous choice has a strictly positive predicted impact on performance relative to EXRA:

$$E[y] = \eta - E[\delta]f + \alpha bJ.$$

The expected perceived stress level of workers is:

$$E[\sigma] = \alpha dJ + [\eta - E[\delta]f + \alpha bJ]\,f.$$

When comparing expected stress levels between ENDO and EXRA two factors need to be considered. First, observing a peer is perceived as stressful. The aggregate importance of this effect for the overall stress level is proportional to the number of workers who observe the peer. Under EXRA the peer is exogenously assigned to half of the population. Under ENDO only workers who desire to observe the peer will do so. If this share of workers is above 50% $(\alpha > \frac{1}{2})$, the first term contributes to an increase in perceived stress in ENDO relative to EXRA. Second, the higher expected performance level under ENDO relative to EXRA (see the discussion above) also contributes to an increase in expected perceived stress, because the second term of the stress function is proportional to performance.

These predictions are fully aligned with the empirical observations that endogenous choice increases both performance and stress relative to EXRA (and RANK).

EXBE: Targeted Matching to the Predicted Most Motivating Option For performance the fact that all workers are "forced" to observe the peer has the following consequences: the workers who desire to observe the peer will be strongly motivated, whereas those who would have preferred not to see the peer will not benefit from this effect. As a

consequence, we end up with the same expression as for ENDO.[55]

$$E[y] = \eta - E[\delta]f + \alpha bJ.$$

The expected perceived stress level of workers is:

$$E[\sigma] = dJ + [\eta - E[\delta]f + \alpha bJ]\, f.$$

The expected level of stress is unambiguously higher than under ENDO (and EXRA). The reason is that more workers observe the peer under EXBE than under both other conditions (as long as $\alpha < 1$) and expected performance is also (at least weakly) higher than in both other conditions.

These predictions are fully aligned with the empirical observation that targeted matching leads to a performance increase comparable in magnitude to that observed under endogenous choice, but results in a substantially larger increase in stress.

D.2.5 Predicted Outcomes for Incentive Treatments

In our incentive treatments (RANK-$, ENDO-$ and EXBE-$), workers get a piece rate p on top of their fixed wage w: $m(y_i) = w + (p + \eta)y_i$. The introduction of a piece rate has the same implications in the model as an increase in the marginal intrinsic utility (i.e. a higher η). The piece rate renders observing a peer more attractive, because it increases the positive marginal effects of higher performance without affecting its downside (i.e. the marginal impact on stress remains constant). As a consequence, the population share of workers willing to observe the peer (α) is expected to increase (see the definition of $\tilde{\delta}$). However, the piece rate has no impact on the directional predictions resulting from our treatment comparisons.

D.2.6 Summary

The following table summarizes the main predictions of this stylized model. For simplicity, we express the effects of the different treatments as changes with respect to the RANK condition (whose predictions also correspond to the predictions of the first round of effort provision in all treatments).

[55] Note that, if instead of assuming that $\epsilon = 0$, we assumed that $\epsilon > 0$ (i.e. if workers who would have preferred not to see the peer still get somewhat more motivated from observing the peer), then expected performance might be slightly higher under EXBE than under ENDO.

Treatment	Δ Peformance	Δ Stress
EXRA	$\frac{\alpha}{2}bJ$	$\frac{dJ}{2} + \left[\frac{\alpha}{2}bJ\right]f$
ENDO	αbJ	$\alpha dJ + [\alpha bJ]f$
EXBE	αbJ	$dJ + [\alpha bJ]f$

Notes: All the changes are indicated relative to the predictions for RANK. The predictions for RANK also correspond to the predictions for the first round of all treatments (the model abstracts from the potentially motivating effects of rank-order information.

The simple example is able to reproduce the main empirical regularities we documented with our experiment. In particular, it makes the following qualitative predictions:

1. Both ENDO and EXBE increase performance relative to EXRA.
2. Performance in EXBE is predicted to be the same as in ENDO.[56]
3. ENDO increases perceived stress relative to EXRA.
4. Perceived stress levels in EXBE are unambiguously higher than in ENDO and EXRA.

Hence, a simple and highly stylized principal-agent model that we augmented to take into account the effects of social comparisons can explain the main findings in our paper.

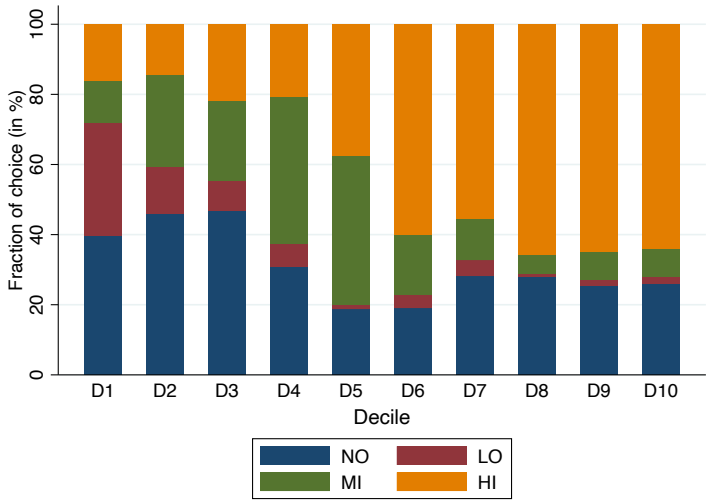
[56]Note that this prediction hinges on the assumption that $\epsilon = 0$. With $\epsilon > 0$, performance in EXBE is predicted to be slightly larger than in ENDO (see also footnote 55).

E Exploring the determinants of choices in ENDO

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

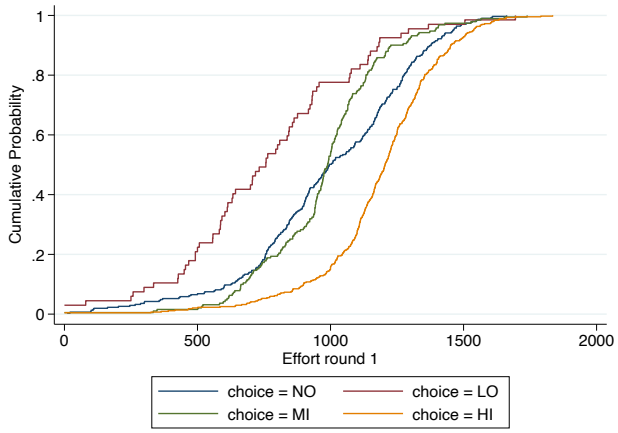
Figure E.2a) depicts the cumulative distribution of period 1 effort, conditional on the cho-

sen 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 E.2b).

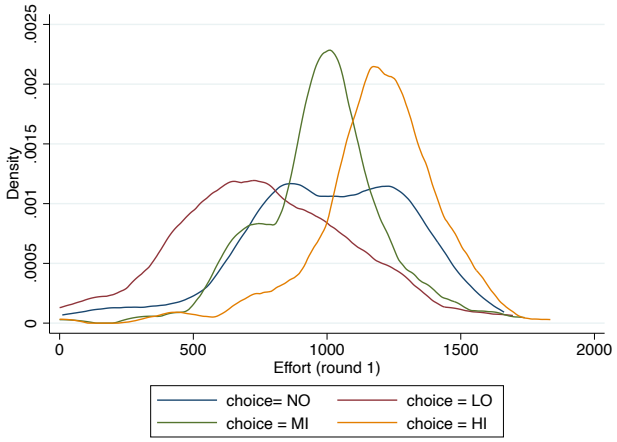
Figure E.3 depicts the aggregate distribution of chosen reference workers, separately by gender. While small differences in proportions exist, the overall choice patterns are qualitatively similar: both gender predominantly prefer to compare to the most productive reference worker; the second largest 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 E.2: Distribution of effort in round 1, conditional on chosen reference worker

(a) cumulative distribution

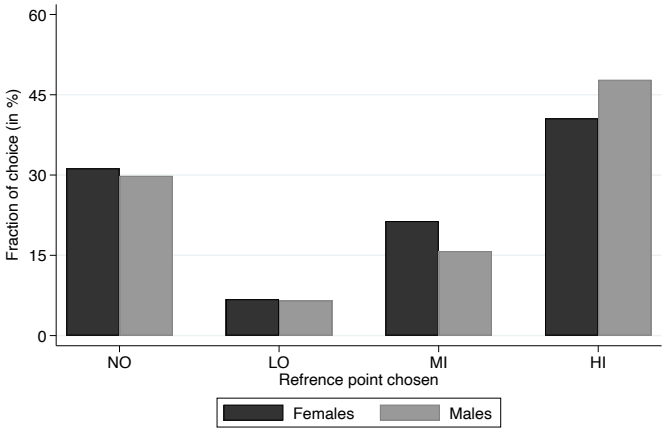


(b) density function



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 E.3: 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.

F Additional analysis: satisfaction and perceived task difficulty

As we discuss in the paper, 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 F.2 and F.3). 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 (Table F.1). 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 Table F.1), with the largest increases in perceived task difficulty being reported in the EXBE condition (see Table F.2 and F.3).

Table F.1: 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 F.2: 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 F.3: 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)
Rank$ 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)
Endo$ 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)
EXBE$ 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

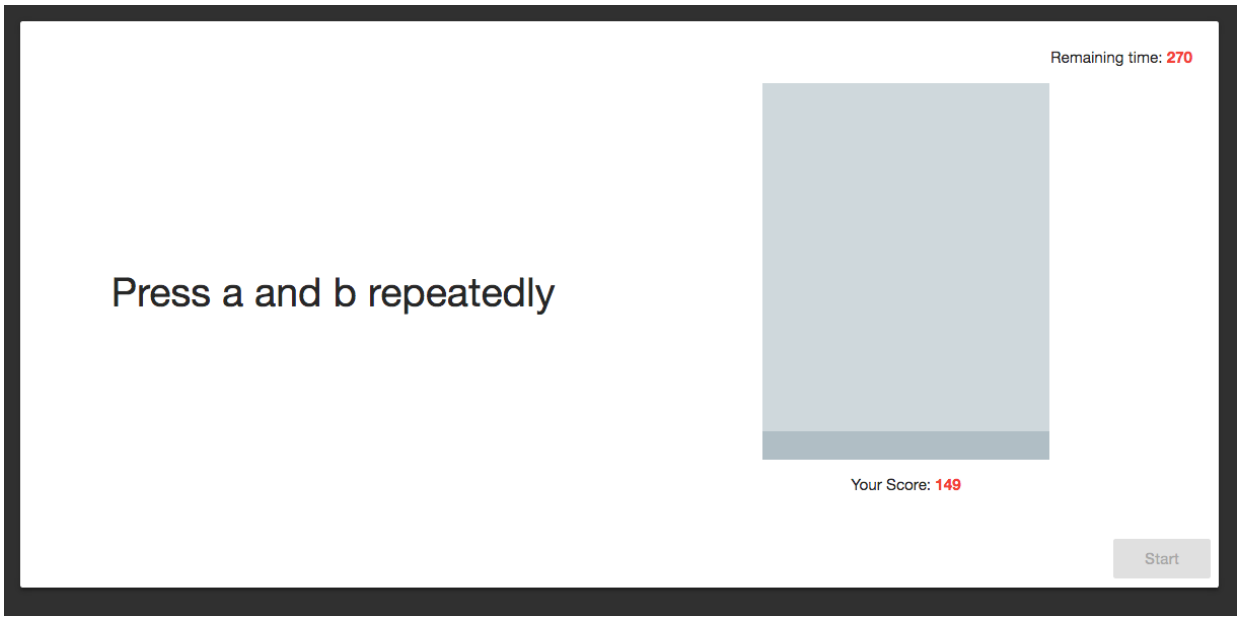
Note: OLS estimations. "Piece-rate (pooled) x P2" is a dummy which equals 1 the treatment that offered a piece-rate in round 2. "Flat wage (pooled) x P2" is participant was in one of the treatment that did not offer a piece-rate in round 2. "RANK x P2" is a dummy which equals 1 if the participant was 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: $*p < 0.1, **p < 0.05, ***p < 0.01$.

G Instructions

G.1 The experiment: EXRA treatment

Figure G.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.

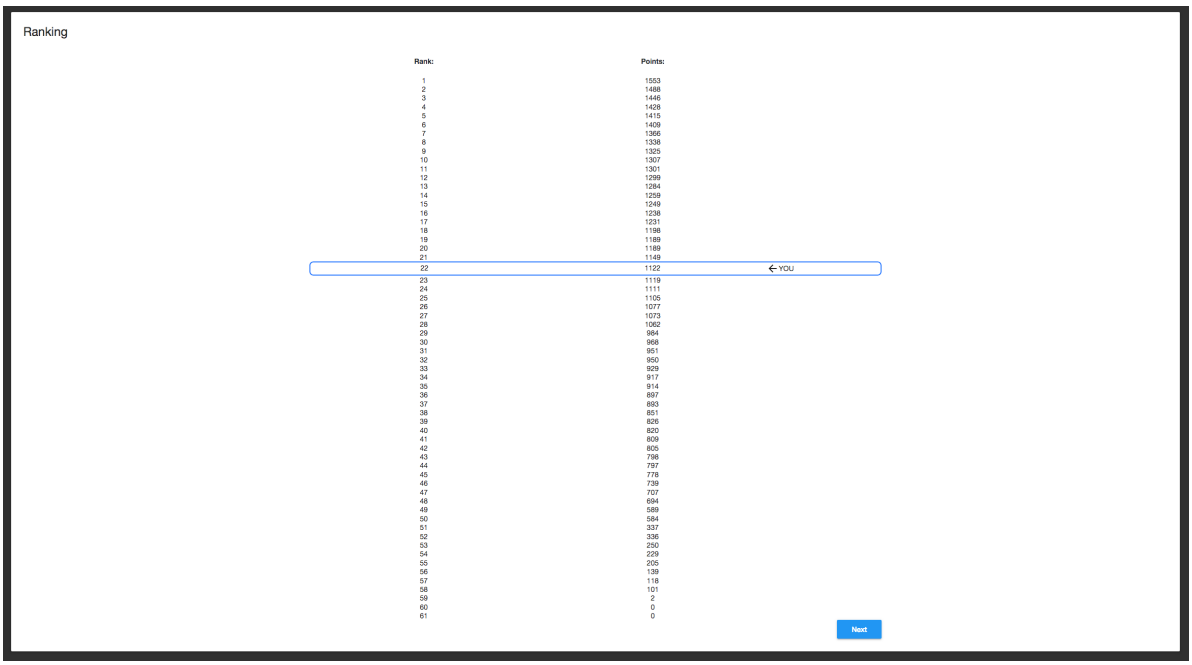
Figure G.1: Real effort task with information about own production only



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.

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 G.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 G.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 G.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.

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 G.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.

G.1.1 The Experiment: 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 G.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 G.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

Figure G.3: Exogenous assignment to a (or no) reference worker. Here, the participant is assigned to the reference worker with average productivity (EXRA-MI, ranked 26).

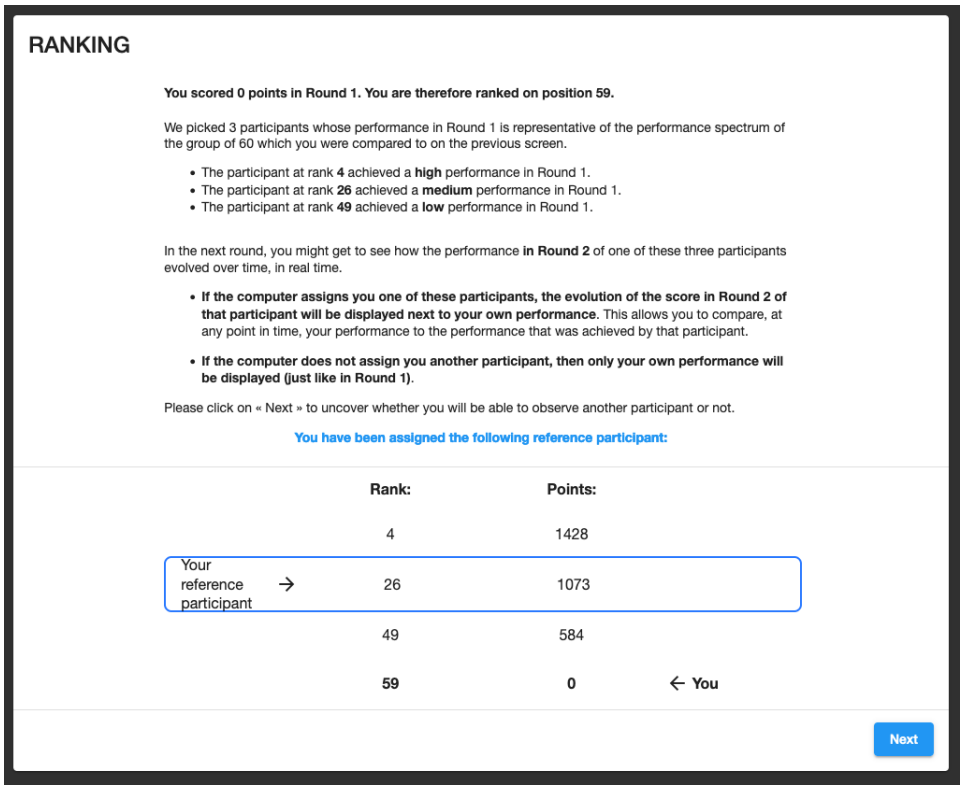
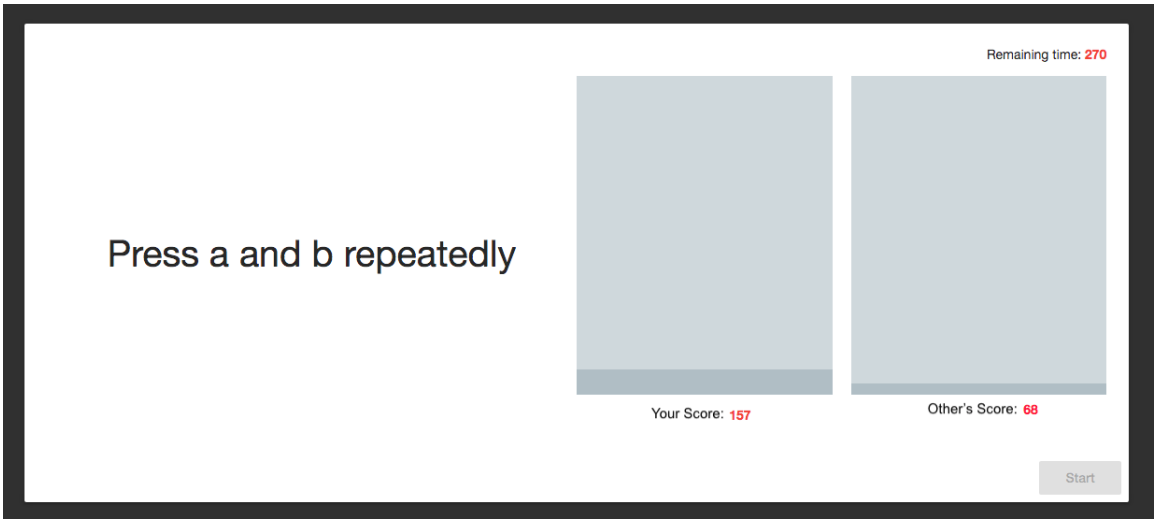
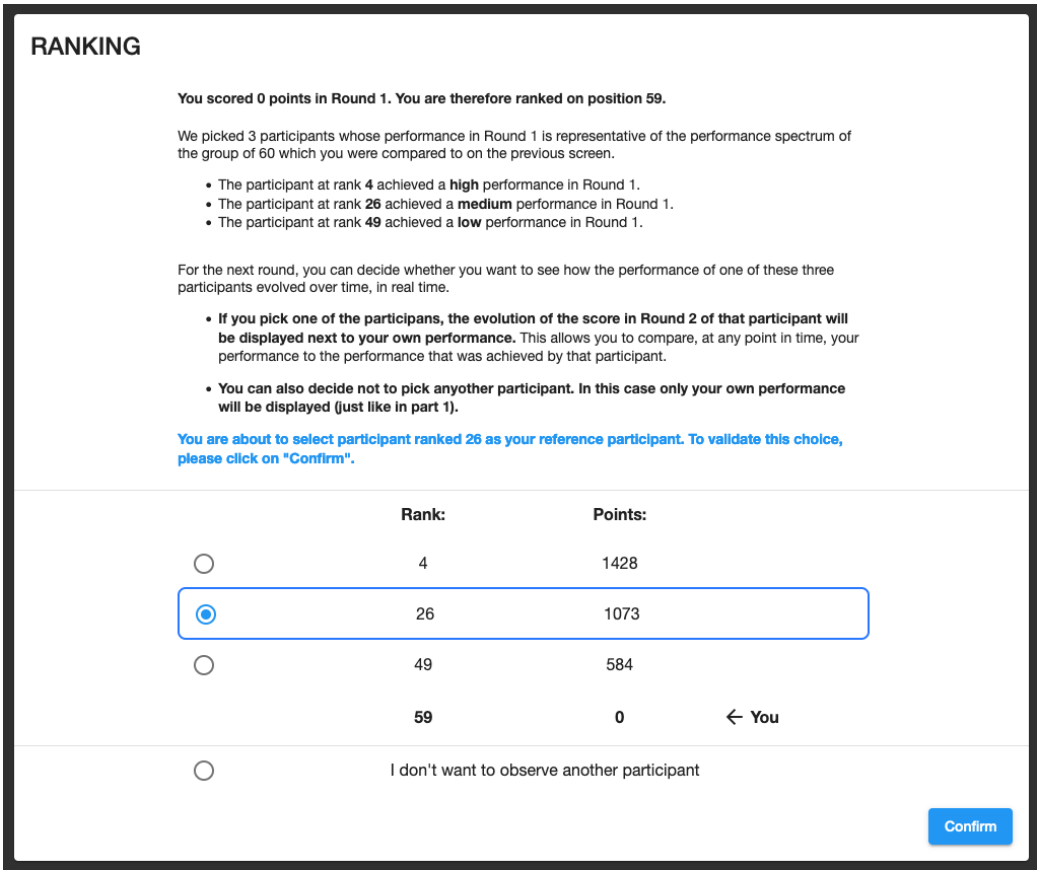


Figure G.4: Real effort task in round 2 during which the worker receives real-time information about the production of a reference worker.



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 G.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.



G.2 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]

G.3 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]

G.4 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 EXRA (and EXBE) 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 EXRA-NO 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]

G.5 Instructions for the follow-up experiment

We provide extensive details on the implementation of this follow-up experiment in Appendix B.7. As discussed therein, the social condition essentially consists of a replication of the EXRA-HI condition. The non-social treatment (PACE-HI) 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 about the same number of points as the reference worker in the EXRA-HI treatment.[57] Just like in our social treatments, the non-social goal in PACE-HI is operationalized as a growing vertical bar.

The first part of this follow-up experiment is identical to the one of the main experiment : participants complete the first period of effort production in isolation (See Supplementary Material G.1 for details on the first round–which is identical across treatments).

At the beginning of round 2, participants in the PACE-HI treatment are informed about the possibility to get assigned to a pacemaker (Figure G.6). Whether or not a worker is assigned to a pacemaker is randomly defined. In the example below, the worker is assigned to the pacemaker and will therefore be able to see it in round 2 (Figure G.7).

[57] To avoid that participants are suspicious, we rounded up the performance of the fast pacemaker to 1600 (instead of 1583 for the highly productive peer). In order not to deceive subjects, we also assign some subjects to a slow pacemaker condition and a condition without pacemaker. Likewise, we assign a few subjects to a randomly assigned reference worker that it *not* the most productive. We discuss this point in Appendix B.7.

Figure G.6: Pacemaker treatment : information about pacemaker

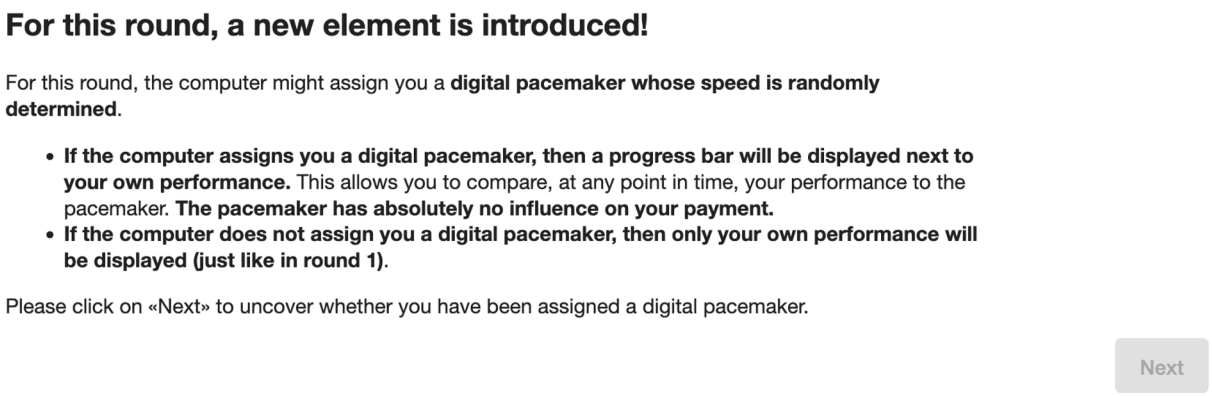
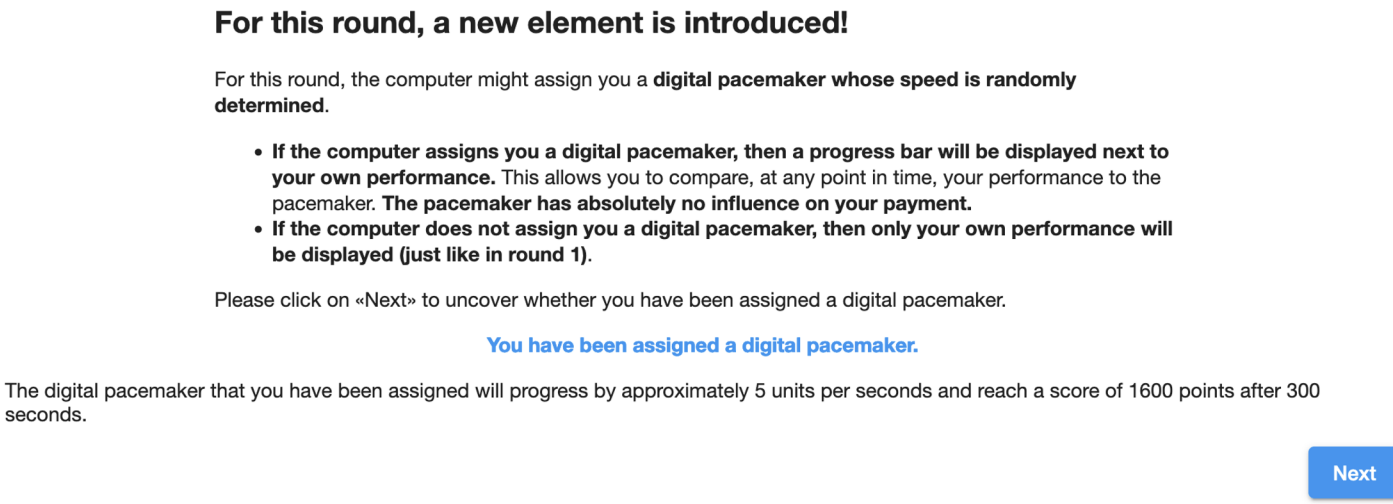


Figure G.7: Pacemaker treatment : revelation of the outcome



If assigned to a pacemaker, workers complete the second round of effort provision while being able to observe the pacemaker (Figure G.8). Like in the treatments with reference workers (i.e., where a social comparison is possible), the pacemaker is displayed as a growing vertical bar. Its speed is set to reach exactly the same number of points as the reference worker used in the social treatment.

Hence, the non-social (pacemaker) treatment is identical to the social treatment, *with the exception that the information provided to participants is non-social*.

Figure G.8: Pacemaker treatment : round 2 in case the participant is assigned the pacemaker