CENTRE FOR DECISION RESEARCH & EXPERIMENTAL ECONOMICS

University of Nottingham
UK | CHINA | MALAYSIA

Discussion Paper No. 2023-03

Patrick Maus, Maria Montero
Martin Sefton

March 2023

# Social reference points and real-effort provision

CeDEx Discussion Paper Series
ISSN 1749 - 3293


CENTRE FOR DECISION RESEARCH & EXPERIMENTAL ECONOMICS

The Centre for Decision Research and Experimental Economics was founded in 2000, and is based in the School of Economics at the University of Nottingham.

The focus for the Centre is research into individual and strategic decision-making using a combination of theoretical and experimental methods. On the theory side, members of the Centre investigate individual choice under uncertainty, cooperative and non-cooperative game theory, as well as theories of psychology, bounded rationality and evolutionary game theory. Members of the Centre have applied experimental methods in the fields of public economics, individual choice under risk and uncertainty, strategic interaction, and the performance of auctions, markets and other economic institutions. Much of the Centre's research involves collaborative projects with researchers from other departments in the UK and overseas.

Please visit http://www.nottingham.ac.uk/cedex for more information about the Centre or contact

Samantha Stapleford-Allen
Centre for Decision Research and Experimental Economics
School of Economics
University of Nottingham
University Park
Nottingham
NG7 2RD
Tel: +44 (0)115 74 86214
Samantha.Stapleford-Allen@nottingham.ac.uk

The full list of CeDEx Discussion Papers is available at

http://www.nottingham.ac.uk/cedex/publications/discussion-papers/index.aspx

# Social reference points and real-effort provision

Patrick Maus[†] Maria Montero[‡] Martin Sefton[§]

March 13, 2023

**Abstract**

We report a laboratory experiment testing whether social reference points impact effort provision. Subjects are randomly assigned the role of worker or peer and the worker observes the peer's earnings before participating in a real-effort task. Between treatments, we exogenously manipulate peer earnings. We find that the workers recall the earnings of their peer and are less satisfied with their own earnings when their peer earns more. Despite this, we do not observe a treatment effect in effort choices. Thus, although our subjects appear to care about income differentials, this does not translate to a change in behavior in our incentivized environment. We relate our results to recent studies of inequality and effort provision.

**Keywords:** social comparisons, reference-dependent preferences, real-effort provision, inequity aversion, relative income concerns

**JEL Codes:** D01, D31, D91, J31

We are grateful to participants at the 2022 CCC conference in Amsterdam and seminars in Nottingham. We also thank Robin Cubitt for very helpful comments. This project was pre-registered at http://osf.io/qvce9. The authors declare that they have no conflict of interest. Ethical approval for the experiments was obtained from the Nottingham School of Economics Research Ethics Committee.

[†]University of Nottingham, United Kingdom. Email: Patrick.Maus@nottingham.ac.uk.
[‡]University of Nottingham, United Kingdom. Email: Maria.Montero@nottingham.ac.uk.
[§]University of Nottingham, United Kingdom. Email: Martin.Sefton@nottingham.ac.uk.

# 1 Introduction

In standard economic models, workers care only about wage levels and their cost of effort when making decisions about their labor supply. However, decisions about labor supply are seldom made in isolation and observing others can be an important source of reference for individuals. For example, the income of others could serve as a social reference point and affect labour supply decisions. Consequently, high social reference points could motivate individuals to change their behavior in order to avoid unfavorable social comparisons. For instance, imagine two identical individuals. Both individuals are equally wealthy but the first one is the richest in his social environment whereas the second one is the poorest. Their income in absolute terms is the same for both individuals but presumably, the second individual is less happy with the status quo and might have stronger incentives to change it.

In particular, social psychology offers a large body of evidence that shows that social comparisons affect the way we feel and behave (Festinger, 1954). People care about their relative status and their decisions are often influenced by what they can observe from others (Bandura, 1977). Social comparisons have been found to be important determinants for our subjective well-being (Ferreri-Carbonell, 2005; Perez-Truglia, 2020), fairness perceptions (Akerlof and Yellen, 1990; Fehr and Gächter, 2000), happiness (Clark and Oswald, 1996; Luttmer, 2005) and health (Marmot, 2005).

Akerlof and Yellen (1990) coined the concept "fair-wage hypothesis". The core assumption of this concept is that people compare their wages to what they perceive as a fair wage and adjust their level of effort accordingly. An important question that remains in this context is what factors determine whether a wage level is perceived as fair or unfair. Austin et al. (1980) show that the comparison of the own income with a peer's income (as a social reference point) has a stronger impact on pay satisfaction than the comparison with an individual (private) reference point like previous earnings. These findings then also raise the question of how individuals would react to advantageous or disadvantageous income comparisons with relevant peers if they have the chance to change their position in the income distribution by modifying their behavior. Genicot and Ray (2017) point out that different exposure to peer outcomes within societies may lead to heterogeneous social reference points which in turn may set higher incentives to work hard for peers from a well-off strata as they are confronted with higher social reference points while peers from a more unfavorable

background have fewer incentives to exert effort. Long-term this could foster equality within groups and inequality between groups.

The main goal of this study is to explore whether and how effort provision is affected by social reference points. Do higher social reference points cause people to supply more effort because they want to avoid unfavorable comparisons? The main challenge to this identification is that social reference points relevant to individuals are usually hard to observe and exogenous variation over social reference points is hard to obtain outside of the lab. We address this research question utilizing an experiment that allows us to exogeneously manipulate peer earnings between treatments.

In our experiment, subjects engage in a piece rate real effort task given the earnings of a peer and they cannot affect their peer's earnings. This is arguably a feature of many work environments where a worker can reduce inequality relative to a social reference point provided by other workers' earnings by working harder, but cannot reduce it by reducing the other workers' earnings. It also provides a clean test of how workers respond to exogenously set social reference points.

In our experiment, only two subjects participate in each session. Each subject is assigned one of the two roles, worker or peer, by a publicly observed coin toss. The peer receives a fixed payment, whereas the worker participates in a tedious real effort task (counting lines). The fixed payment to the peer is varied across two treatments. Consequently, any difference in the average level of effort provision by the workers between treatments can be traced back to social reference points. In our experiment, no worker can observe or be observed by other participants. This minimizes other peer effects, such as peer pressure, imitation, or learning (Sacerdote, 2001; Falk and Ichino, 2006; Mas and Moretti, 2009).

To derive the main hypothesis that our experiment is designed to test, we formalize a simple Fehr-Schmidt model of inequality aversion around the peer's earnings. The manipulation of the peer's earnings changes the incentives for the workers that want to avoid unfavorable social comparisons. Consequently, our main hypothesis is that workers provide more effort on average when the earnings of their peer are relatively high (*High treatment*) compared to when they are relatively low (*Low treatment*).

In our experiment, we find that workers remember their peers' earnings and appear to care about them. The workers are significantly less happy with their earnings when their peer earns £7.10 (*High treatment*) compared to when it earns £2.90 (*Low treatment*). While workers are less happy when their respective

peer earns more, this does not translate to a change in working behavior in the real-effort task. The difference in effort provision between treatments is very small and not statistically significant. Together, this suggests that people care about income differences but this does not necessarily translate to a change in behavior in incentivized environments.

This experiment was pre-registered, with sample sizes determined through a power analysis used to ensure adequate statistical power. The power analysis was based on assumptions about effort provision using data from a previous experiment (Gagnon et al., 2020). A surprising result from our experiment was that efforts, in both treatments, substantially exceeded our expectations and were inconsistent with our assumptions. Indeed, many workers provided maximal effort, even though our experiment was designed so that workers would choose when to stop working and the maximal effort constraint would not bind. A consequence is that similar efforts across treatments may be due to a "ceiling effect" whereby real effort costs were so low as to make effort insensitive to incentives.

Because the effort provided by our subjects in our pre-registered experiment was substantially greater than anticipated and many people solved the maximum amount of lines, we conducted a follow-up experiment with a modified and more difficult working task. We do this by raising workers' effort costs to ensure that in our experiment enough participants could be influenced by concerns of behindness aversion. Our re-calibration successfully reduced effort levels such that no subjects exerted maximal effort. The results of our re-calibrated experiment are qualitatively very similar to the original experiment and again we find no indication that social reference points affect real-effort provision.

The paper is organized as follows. Section 2 discusses the relevant literature. Section 3 explains the experimental design in detail. Section 4 discusses the behavioral predictions and derives our main hypothesis. In Section 5 we present the result of our original experiment. Section 6 discusses and presents the results of our re-calibrated experiment. Section 7 concludes.

## 2 Related Literature

This study aims to contribute to the strands of literature on reference dependence, peer effects, and wage inequalities. Firstly, the study is linked with the literature on reference dependence. So far, a lot of this literature has focused

on individual decision making where reference points are based on expectations or the status quo (see e.g. Kőszegi and Rabin (2006)). For instance, Gill and Prowse (2012) investigate whether individuals are disappointment averse when they compete in a real effort sequential-move tournament. They find that the second mover shies away from working hard when she observes that the first mover has worked hard, and tends to work relatively hard when she observes that her competitor has put in a low effort which is consistent with an expectation-based model of disappointment aversion. Gächter et al. (2018) compare "social" and "asocial" versions of the Gill and Prowse experiment, where the scope for social comparison is removed in the latter. They find behavior in social and asocial treatments to be similar, suggesting that social comparisons have little impact in this setting.

Abeler et al. (2011) also provide evidence that expectations-based reference points matter in the context of effort provision. In their experiment, subjects work on a tedious and repetitive task. After each repetition, they decide whether to continue or to stop working. They get a piece rate, but receive their accumulated piece-rate earnings only with a 50 percent probability, whereas with a 50 percent probability they receive a fixed, known payment instead. Which payment subjects receive is determined only after they have made their choice about when to stop working. The only treatment manipulation is a variation in the amount of the fixed payment. In their experiment workers worked substantially longer and earned more money if their expectations were high compared to low expectations. While there is some evidence that expectations-based reference points matter when it comes to effort provision, evidence is scarce when it comes to different sources of reference points. We complement the literature of reference points and effort provision by using social outcomes as reference points while our design allows us to keep expectations constant between treatments.

More recently, Schwerter (2023) investigates experimentally whether social reference points impact individual risk-taking. Decision-makers in his experiment observe the earnings of a peer subject before making a risky choice. This allows him to manipulate the peer earnings across two treatments exogenously. He finds a significant treatment effect on risk-taking, i.e. decision-makers make more risk-seeking choices if the earnings of their peer are relatively large. The experimental design of Schwerter (2023) is very similar to ours where people make effort choices but prior observe peer earnings.

Secondly, our study contributes to the literature on peer effects. The most closely related studies to ours investigate the impact of peers on performance.

Card et al. (2012) find that workers with salaries below the median for their pay unit and occupation report lower pay and job satisfaction, while those earning above the median report no higher satisfaction. Likewise, below-median earners are significantly more likely to look for a new job, while above-median earners are unaffected. Their findings suggest that job satisfaction depends directly on relative pay comparisons.

Falk and Ichino (2006) recruited subjects to fill letters into envelopes with a remuneration independent of their performance. In the first treatment, subjects worked alone in a room, and consequently peer effects are ruled out. In the second treatment, two subjects work in one room at the same point in time which makes peer effects possible. They find that peer effects raise productivity and that low-productivity workers are the most sensitive to the behavior of peers. Mas and Moretti (2009) find strong evidence of positive productivity spillovers from the introduction of highly productive personnel into a shift using high-frequency data on worker productivity from a supermarket chain. The effort of the workers is positively correlated to the productivity of workers who see them but not workers who do not see them. In all these studies, individuals usually observe the results and behavior of others and expect others to observe their behavior and results. As a consequence, this does not allow them to distinguish between relative concerns and additional peer effects, such as social pressure, imitation, and learning. Our study aims to test a specific channel through which peer effects might work, namely that people compare their income to that of a peer and adjust their behavior accordingly.

Thirdly, our study is also related to the experimental literature that investigates how wage inequalities influence effort provision. Gächter and Thöni (2010) find that disadvantageous wage discrimination leads to lower levels of effort while advantageous wage discrimination does not increase the levels of effort on average using a three-person gift-exchange game. In a similar vein, Cohn et al. (2014) conducted a field experiment to examine how social comparison affects workers' effort provision if their wage and/or the wage of their co-worker is cut. Workers were assigned to groups of two, performed the same task, and received the same performance-independent hourly wage. They find that workers decrease their level of effort less in consequence of a pay cut if not only their wage but also the wage of their co-workers is cut. However, in the experimental setting of Gächter and Thöni (2010) and Cohn et al. (2014) the wage of the employees does not depend on performance and their behavior could also be influenced by reciprocal motives. In the environment of labor supply

that we study here the performance of a subject is directly linked to its earnings and there are no reciprocal relationships.

More closely related to our study, Bracha et al. (2015) test the hypothesis that low wage levels compared to others decrease labor supply. For testing this, they offer participants to choose how long they want to work on a real-effort task for a piece-rate pay level that is either high or low. In one treatment, the subjects are only aware of one pay rate, whereas in the second treatment, they have a natural reference point - the other piece-rate pay level offered. They find that relative pay comparisons affect labor supply, such that lower-paid individuals supplied significantly less work time relative to higher-paid individuals and significantly less time than when they were unaware of the higher piece rates.

Note that, Bracha et al. (2015) do not implement a social reference point that has pre-determined reference earnings associated with it. The subjects are aware of the different piece-rate payments of their peers (and consequently might form beliefs about the performance of their counterparts) but they do not know the absolute income of their peers. The absence of a deterministic social reference point might modify working behavior. For instance, the earnings of the individual's peer are not salient and could be strongly correlated with the beliefs about its peer's working motivation. Moreover, the absence of a deterministic social reference point makes it impossible to know whether a decision-maker could avoid a disadvantageous social comparison by providing additional effort.

Gagnon et al. (2020) investigate the effect of neutral and gender-discriminatory unfair chances on real-effort provision. In their experiment, workers engage in a real-effort task for a piece-rate wage on an online labor platform. They randomize workers into treatments where they control relative pay and chances to receive a low or a high wage. They find that unequal pay affects the labor supply of discriminated workers but does not change the effort-provision of the high-wage workers significantly, irrespective of whether the low wage is the result of fair or unfair chances. However, similar to Bracha et al. (2015) they do not introduce a deterministic social reference point. In contrast to the experiments of Bracha et al. (2015) and Gagnon et al. (2020), we do not manipulate the piece-rate incentives between two workers performing an identical task. In our setting, one individual is endowed with a fixed payment without performing a task and consequently, this payment might serve as a social reference point for a worker who is rewarded according to the same piece-rate wage between treatments.

# 3 Experimental Design

The experiment is designed to create an environment that allows for a precise measurement of effort provision and in which we can exogenously influence a reference point. Between the two treatments, we manipulate predetermined peer earnings exogenously. This allows us to test the influence of a social reference point on effort provision by comparing the behavior of the subjects between treatments. Our study is preregistered at the Open Science Framework (Doi: 10.17605/OSF.IO/QVCE9).

The experiment was conducted in virtual rooms in Microsoft Teams with students of the University of Nottingham in June 2021. Each experimental session consisted of two subjects only. The subjects were recruited via Orsee (Greiner, 2015). The experimental materials are reproduced in Appendix B to Appendix E.

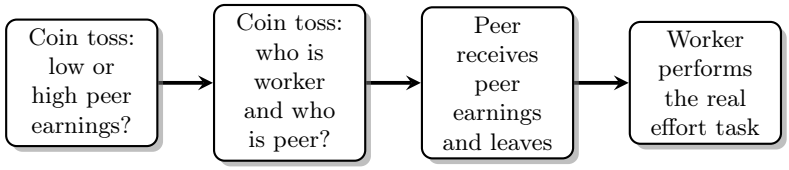



Figure 1: Procedural overview of the experiment

Upon the arrival of the subjects in the virtual room, the subjects are instructed that only one of the two subjects will participate in a working task and that the second subject will receive a lump-sum payment and has to leave the room. First, the experimenter tosses a virtual coin to determine the size of the lump-sum payment. In this way, subjects are randomly assigned to either a *High treatment* (lump-sum payment = £7.10) or *Low treatment* (lump-sum payment = £2.90). After both subjects observe the coin toss and are informed about the size of the lump-sum payment, the experimenter tosses another coin to determine which subject participates in the working task and which subject has to leave with the lump-sum payment. In the experiment we referred to these roles as Player A and Player B, for the remainder of the paper we will refer to them as worker and peer.

We randomize the size of the lump-sum payment to the peer in each experimental session because we want to ensure that the lump sum payment does not convey different information about earnings to be expected from the working

task. If the lump-sum payment conveys any information in our experiment, this information would be conveyed by the average of the two possible lump-sum payments which is constant across treatments.

As the working task, we chose the same task as used in Gagnon et al. (2020) where workers are asked to copy lines into a textbox. Each worker receives £0.06 per correctly entered line. In addition, workers received a fixed payment of £2. This task does not require any prior knowledge and the performance of the worker is easily measurable. The task is clearly artificial and it should be obvious that the performance of a subject is of no value to the experimenter. Consequently, this minimizes any tendency for workers to use effort in the experiment to reciprocate the experimenter.

The task gets increasingly harder over time, such that the length of the lines increases with the number of completed lines whereas the piece-rate remains unchanged. Consequently, engaging in the task should also become less attractive for the workers over time. The workers are instructed that they can work at most for 90 minutes but they can stop working at any point in time. This is identical to Gagnon et al. (2020) with the exception that workers have slightly more time and we introduced a higher maximum number of solved lines. We do this to allow the majority of subjects to work as much as they want. The workers were unaware that there is a maximum number of lines they can solve during the experiment. We use the number of lines entered correctly as the measure of effort provision. As the subject works in isolation, other sources of peer effects, other than earnings comparisons, are mitigated (e.g. imitation, learning, or peer pressure). Then, the worker answers a short questionnaire and the experiment is over.

The questionnaire includes questions about whether the subject remembers the size of the lump-sum payment, regarding their gender, the gender of their peer, the perceived social closeness to their peer (Gächter et al., 2015), and self-reported competitiveness. Moreover, we elicit (without incentives) parameters of a Fehr-Schmidt inequality aversion model following Blanco et al. (2011). To elicit the aversion against disadvantageous equality $\alpha_i$ we use responder deci-

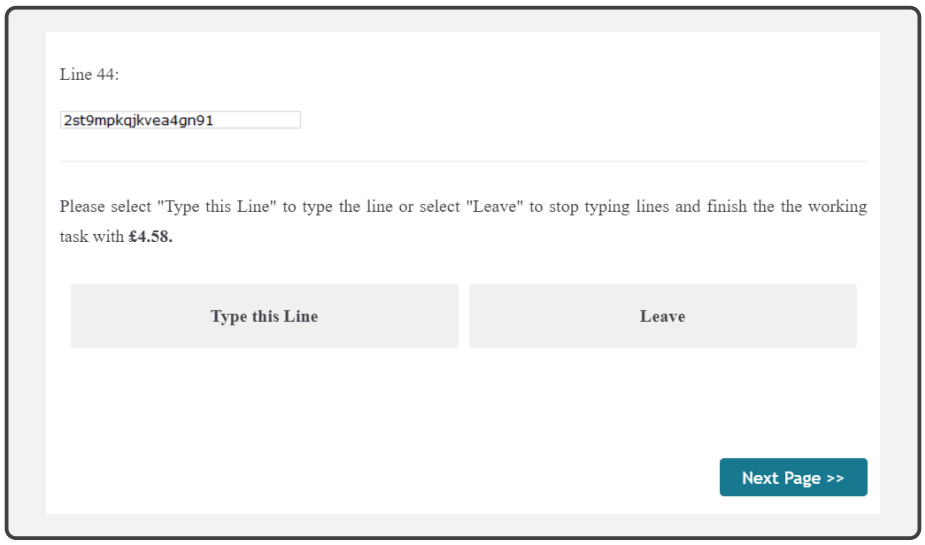



Figure 2: Screenshot of the working task

sions from an ultimatum game.[1] For eliciting the aversion against advantageous inequality $\beta_i$ we use a modified dictator game.[2] We use non-incentivized measures of $\alpha_i$ and $\beta_i$ to ensure that the only source of payoff comparisons are the earnings from the work task.

[1] Suppose $x_i'$ is the lowest offer the responder accepts and $x_i'-1$ is the highest offer the responder rejects, where offers are integers between 0 and 10 out of a total amount of 20. Then a responder with well-behaved preferences is indifferent between accepting some offer $x_i \in [x_i'-1, x_i']$ and both players receiving nothing. From $U_i(x_i, 20-x_i) = x_i - \alpha_i(20-x_i-x_i) = 0$ we can determine the point of indifference $\alpha_i = \frac{x_i}{2(10-x_i)}$. Following Blanco et al. (2011) we approximate $x_i$ as the average of $x_i'$ and $x_i'-1$ when there is no more than one switch-point in the responder's strategy. The workers who accept all offers are assigned $\alpha_i = 0$ and the workers who reject all offers are assigned $\alpha_i = 4.5$.

[2] We obtain $\beta_i$ by finding the egalitarian allocation $(x_i,\ x_i)$ that makes the dictator indifferent between keeping the entire endowment $(10,\ 0)$ and $(x_i,\ x_i)$. Suppose the dictator switches to the egalitarian outcome at $(x_i',\ x_i')$. That is, he prefers $(10, 0)$ over $(x_i'-1,\ x_i'-1)$ but $(x_i',\ x_i')$ over $(10, 0)$. We conclude that he is indifferent between $(10, 0)$ and the $(x_i, x_i)$ egalitarian outcome where $x_i \in [x_i'-1, x_i']$ and $x_i' \in \{1, \ldots, 10\}$. Thus, $\beta_i = 1 - \frac{x_i}{10}$. Again, we approximate $x_i$ as the average of $x_i'$ and $x_i'-1$. The workers who always choose the egalitarian outcome are assigned $\beta_i = 1$ and the workers who never choose the egalitarian outcome are assigned $\beta_i = 0$.

# 4 Theoretical Predictions & Hypotheses

## 4.1 Theoretical Predictions

Consider two individuals, worker $i$ and its peer $j$. Worker $i$ is engaged in a real-effort task receiving a show-up fee $R$ and piece-rate wages $we$, whereas its peer $j$ receives a lump-sum payment $F$. Worker $i$ chooses her optimal effort level $e^*$ given her wage $w$, the social reference point $F$ and her cost function $c(e;\lambda_i)$ including an individual-specific cost parameter $\lambda_i$. For simplicity, we assume the cost function to be quadratic. Worker $i$ chooses her optimal effort level to maximize her utility function:

$$U_i = R + we_i - \alpha_i \max\{(F - R - we_i, 0)\} - \beta_i \max\{(R + we_i - F, 0)\} - \frac{\lambda_i e_i^2}{2} \quad (1)$$

The first and second term on the RHS of equation (1) corresponds to the utility of monetary earnings derived from working, the third and fourth term account for a possible disutility in consequence of payoff inequalities and the final term is the utility cost of providing effort. The modeling of disutility created by a payoff inequality follows Fehr and Schmidt (1999). The third term of the RHS measures the disutility from disadvantageous payoff inequality and the fourth term measures the disutility from advantageous payoff inequality, with $\alpha_i \geq \beta_i \geq 0$. Consequently (1) collapses to:

$$U_i = \begin{cases} R + we_i - \alpha_i(F - R - we_i) - \frac{\lambda_i e_i^2}{2} & \text{if } we_i < F - R \\ R + we_i - \beta_i(R + we_i - F) - \frac{\lambda_i e_i^2}{2} & \text{if } we_i \geq F - R \end{cases} \quad (2)$$

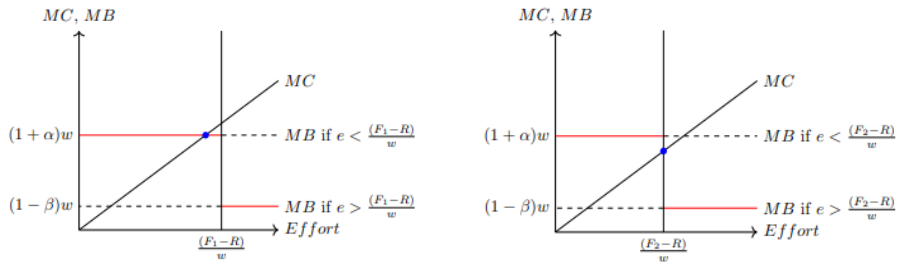
Differentiating (2), we obtain the following FOCs and optimal levels of effort:

$$\frac{\partial U_i}{\partial e} = \begin{cases} w + \alpha_i w - \lambda_i e_i & \text{if } we_i < F - R \\ w - \beta_i w - \lambda_i e_i & \text{if } we_i \geq F - R \end{cases} \quad (3)$$

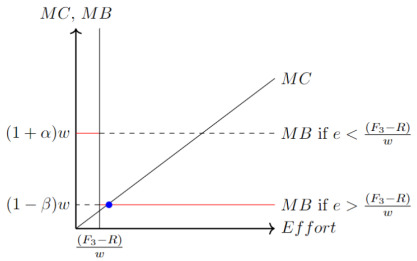
$$e_i^* = \begin{cases} \frac{(1+\alpha_i)w}{\lambda_i} & \text{if } \frac{(1+\alpha_i)w^2}{\lambda_i} \leq F - R \\ \frac{F-R}{w} & \text{if } \frac{(1-\beta_i)w^2}{\lambda_i} < F - R < \frac{(1+\alpha_i)w^2}{\lambda_i} \\ \frac{(1-\beta_i)w}{\lambda_i} & \text{if } \frac{(1-\beta_i)w^2}{\lambda_i} \geq F - R \end{cases} \quad (4)$$

The marginal utility in (3) can be viewed as the difference between marginal benefit and marginal cost of entering lines, where the marginal benefit depends on whether earnings exceed or fall short of the reference earnings. The optimal

effort in (4) consequently depends on whether the reference earnings are comparatively large or small. This is in contrast to the standard model where the optimal level is $e_i^* = \frac{w}{\lambda_i}$ independently of $F$.




(a) $\frac{(1+\alpha)w^2}{\lambda} \leq F - R$ (b) $\frac{(1-\beta)w^2}{\lambda} < F - R < \frac{(1+\alpha)w^2}{\lambda}$




(c) $\frac{(1-\beta)w^2}{\lambda} \geq F - R$

Figure 3: Optimal individual effort for a given $w$, $R$, $\lambda$, $\alpha$ and $\beta$ when the reference earnings are (a) high, (b) medium or (c) low.

Figure 3 graphically illustrates the three cases to consider for a given $w$, $R$, $\lambda$, $\alpha$ and $\beta$. The reference earnings can be either (a) high, (b) medium or (c) low. Panel (a) describes the case where the individual faces high reference earnings. Here, she chooses the level of effort that corresponds to the intersection between her marginal cost function and her marginal benefit function which yields earnings below the reference earnings. Consequently, if the reference earnings are shifted further to the right, this does not alter her optimal level of effort provision. Panel (b) describes the case where the individual faces an intermediate social reference point. This case illustrates the situation where she chooses her level of effort such that she matches the earnings of her peer. Consequently, if the reference point is shifted further to the right, she increases her level of effort such that it corresponds to the earnings of the new social reference point or up to the point when the marginal cost function and marginal

benefit function intersect. Panel (c) describes the case where the individual faces a low social reference point. If the reference point is shifted to the right this might have three possible different effects. If the shift is only very slight, the optimal level of effort does not change. If the shift is moderate, she increases her level of effort such that she matches the reference earnings. If the shift is large, she increases her effort level to the point where the marginal cost function and marginal benefit function intersect. Consequently, an increase in the social reference point leads to at least a weakly higher level of effort in all three panels.

**Proposition 1.** *Higher social reference points lead to weakly higher levels of effort.*

Note that the assumption that $\beta_i > 0$ can be relaxed. Our predictions also hold, as long as we assume $\alpha_i > |\beta_i|$. Under this condition, the marginal benefit function $(1 + \alpha_i)w$ is always above the marginal benefit function $(1 - \beta_i)w$ and consequently intersects the marginal cost function at a higher level of effort.

## 4.2 Hypotheses

Given our theoretical analysis, we arrive at the following testable hypotheses for the comparison between the two treatments:

**Main Hypothesis**: The distribution of effort in the "high reference point treatment" stochastically dominates the distribution of effort in the "low reference point treatment".

We test this hypothesis using a one-sided Wilcoxon-rank sum test. To determine the appropriate sample size we conducted a power analysis. Based on the results of this we determined that for our experimental parameters a sample of 180 worker-peer pairs, randomly assigned to *High treatment* and *Low treatment* with equal probability would provide adequate power ($> 80\%$) of the test (see Appendix A for details).

According to our theoretical analysis, the optimal effort level can be determined by two processes. Either the individual chooses the effort level that corresponds to the intersection between the marginal benefit function and the marginal cost function, or she matches the reference earnings. Assuming that we have cases where individuals choose the effort level that corresponds to the reference earnings, our model suggests not only that the effort levels differ between treatments, but they differ in a very specific way. Consequently, our subsidiary hypothesis predicts a higher probability of matching the earnings of

the high reference point in the high reference point treatment compared to the low reference point treatment.

**Subsidiary Hypothesis:** The probability for the workers to stop at earnings = size of the high reference point is higher in the "high reference point treatment" compared to the "low reference point treatment".

While our theory also predicts a higher likelihood for workers to stop at the low reference point in the low reference point treatment compared to the high reference point treatment, we do not formalize this as a testable hypothesis. The reason is that, based on our experimental parameters and assumptions about the distribution of effort cost and inequality aversion parameters, we do not have sufficient power to test this prediction (see Appendix A for details).

From the perspective of expected utility theory as well as models of status-quo-based (Tversky and Kahneman, 1979) and expectations-based (Kőszegi and Rabin, 2006) reference points, no treatment effect on effort provision is predicted since decision-makers faced the same working task across treatments and knew that their earnings were determined solely by their performance. Consequently, we are confident that a difference in effort between the "high reference point treatment" and the "low reference point treatment" can be solely attributed to the effect of peer earnings.

# 5 Results

## 5.1 Main Analysis

In total, 360 subjects participated in the experiment. 180 of whom participated in the work task and 180 of whom were allocated the role of a passive peer. The random assignment to treatments resulted in 102 workers in the *High treatment* and 78 workers in the *Low treatment*. The workers in both treatments are well balanced. There are no significant differences between the two treatments in terms of gender (54.9% in *Low*, 55.1% in *High*). This resulted in 53% of workers paired with a peer of the same gender (52.9% in *Low*, 53.8% in *High*). Further, the level of competitiveness elicited in the post-experimental questionnaire is very similar (4.92 in *Low*, 5.06 in *High*). All sessions were conducted online with students of the University of Nottingham in June 2021. As we preregistered, we use one-sided tests for our hypotheses and two-sided tests for the more exploratory analysis.

It appears that workers cared about the earnings of their peer. The large majority of the workers could remember the earnings of their peer very precisely. Further and especially interesting, the workers were significantly happier when they were in the *Low treatment* treatment compared to the *High treatment* (4.40 in *Low*, 3.87 in *High*, two-sided Wilcoxon rank-sum test, p-value= 0.030). This shows that workers cared about the income of their peers on a psychological level. However, this did not translate to a change in working behavior in our experimental task.

We do not find support for our main hypothesis in the original experiment. In the *Low treatment* with peer earnings of £2.90, the workers solved on average 110.34 lines correctly. In the *High treatment* with peer earnings of £7.10, the workers solved on average 112.21 lines. The treatment difference of 1.78 correctly solved lines is negligible and corresponds to a marginal effect of 1.02%. To test for the equality of effort provision we use the one-sided Wilcoxon rank-sum test yielding a p-value of 0.495.[3]

**Result 1:** The workers do not work significantly harder in the *High treatment* compared to the *Low treatment* in the original experiment.

What makes the absence of a treatment effect in terms of effort provision especially interesting is that workers were less happy with their earnings in the *High treatment* compared to the *Low treatment*. The treatment difference in terms of happiness remains significant once we control for the individual performance in the working task in an OLS regression (p-value $< 0.01$). Together, this suggests that the workers cared about the earnings of their respective peer but did not change their behavior accordingly. A possible reason why we do not observe a treatment effect in effort provision could be that some subjects reduce their effort when they observe a high reference point. Then, we would expect a greater variance in effort provision in the *High treatment* compared to the *Low treatment* which we do not find (two-sided Kolmogorov–Smirnov test, p-value=0.844).

The modal choice in both treatments is to solve the maximum number of 170 lines. Based on results in Gagnon et al. (2020) we expected the participants to exert considerably less effort and did not expect the ceiling to be binding.This might indicate that the working task has been perceived as too easy and not

[3]In our analysis, we give priority to non-parametric tests because they do not assume that error terms are normally distributed.

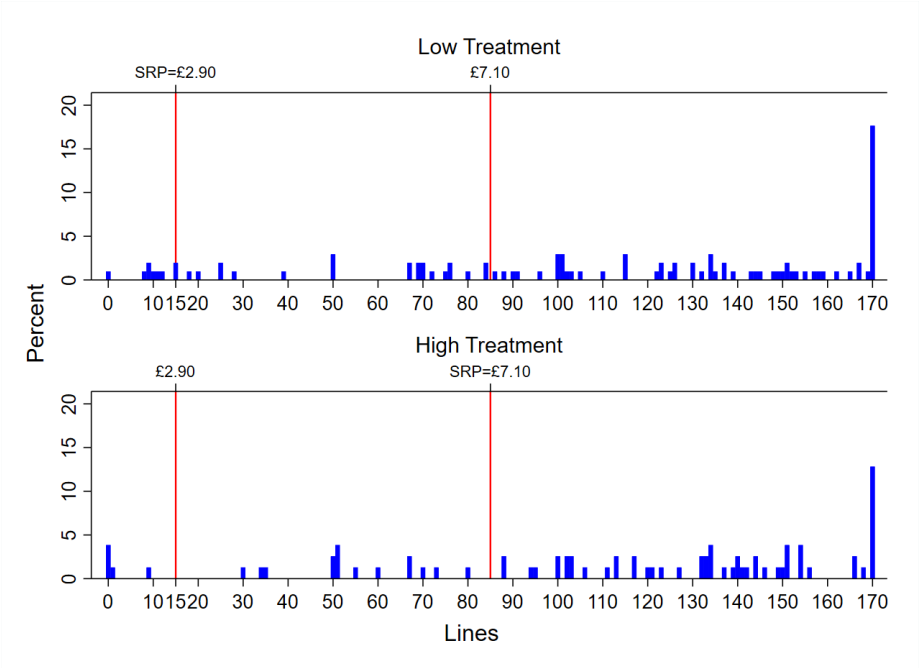



Figure 4: Histograms of correctly solved lines by treatment where the social reference points are indicated by the red lines in the original experiment

tedious enough in our experimental setting. It could be that the cost of effort is not only determined by the number of lines solved but also by the actual time workers spend on the working task. We, therefore, consider the time spent working as an alternative measure of effort provision. The average times working in both treatments do also not differ significantly from each other. The workers in the *High treatment* work on average for 53.92 minutes, while the subjects in the *Low treatment* work on average for 52.98 minutes (Wilcoxon rank-sum test, p-value= 0.474).

As shown in Section 4, our model does not only predict treatments to be different but to be different in a very specific way. Our subsidiary hypothesis predicts a higher probability of stopping when the accumulated earnings equal the peer's earnings in the *High treatment*.

Figure 4 shows a histogram of correctly solved lines for each treatment (*Low treatment* in the top panel, *High treatment* in the bottom panel). First of all, one can see that stopping decisions are dispersed over a wide range. Some workers stop directly or very early, others solve all 170 lines. This heterogeneity is

what one would expect given that productivity, cost of effort, and opportunity cost differ across workers. We are interested to see whether there are systematic differences in the clustering between treatments in terms of clustering of stopping decisions exactly at the peer's earnings: In neither treatment does a subject stop at exactly 85 lines, which corresponds to the income of the peer in the *High treatment*. This does not allow us to test our hypothesis using our preregistered test. If we compare the number of workers stopping in the range between 80 and 90 correctly solved lines, we find no significant differences between treatments (one-sided Fisher's exact test, p-value= 0.474). This also holds for the ranges between 75 and 95 (one-sided Fisher's exact test, p-value= 0.258) and 70 and 100 correctly solved lines (one-sided Fisher's exact test, p-value= 0.260).

**Result 2:** The probability for the workers to stop at earnings = size of the high reference points is not higher in the *High treatment* compared to the *Low treatment* in the original experiment.

## 5.2 Discussion of Main Results

Since we do not find a significant difference regarding the levels of effort provision between treatments, we use additional data from our questionnaire to check whether our design induced a salient reference point. To do this, we investigate whether workers remembered the reference point while working on the task. Finally, we discuss whether our results could be related to the degree of behindness aversion and leading aversion of our subject pool.

### 5.2.1 Salience

Given that we found no differences in effort provision between the two treatments, it is an important question to address whether the workers remembered the reference point while they did the task. In our questionnaire, we ask the participants who took part in the work task whether they can remember the amount of the payment we made to the other participant. 75% (73.5% in *Low*, 77% in *High* of all workers remember the exact size of the reference point). If we allow for a deviation of $\pm$ 1 pound, 97.8% of all workers answer correctly. In summary, this means that the workers can remember their peer's income fairly well, but it does not seem to influence their decision to work in a significant way. We also ask the workers in our questionnaire what factors influenced their decision to work. 26.1% of all workers mention their peer's income as an influencing

factor for their effort provision. Interestingly, workers are significantly more likely to mention their peer earnings in the *High treatment* (37.1%) compared to the *Low treatment* (17.6%). We do not find a statistically significant difference in effort provision between treatments if we restrict the sample to workers that mentioned their peer earnings as a relevant factor (Wilcoxon rank-sum test, p-value= 0.637). Together with the finding that workers are significantly less satisfied with their earnings when their peer earns relatively more, this yields suggestive evidence that the treatment manipulation itself worked well but did not lead to significant changes in the effort provision of the workers.

### 5.2.2 Behindness Aversion

In our questionnaire, we elicited a measurement of each subject's level of behindness aversion. One possible reason why our treatment manipulation does not lead to the expected results is that a large fraction of our workers does not care about the income of others which would be reflected in very low levels of behindness aversion. The alphas we elicited in our experiment are on average 0.71. This is very much in line with the literature. For instance, Beranek et al. (2015) elicited alphas and betas at the same university in an incentivized way. Comparing the distribution of alphas between the two studies, we find that workers in our experiment are slightly more likely to report smaller values of alpha. In our experiment 63.3% of the workers report an alpha below 0.4, whereas in their experiment 54% do so but overall the distributions are very similar (mean in our experiment = 0.71, mean in Beranek et al. (2015) = 0.75). Originally, we expected that workers with a higher degree of behindness aversion would respond more strongly to the treatment manipulation than people with a lower degree. We can find no support for this in our experimental data. We find no significant relationship between the degree of behindness aversion and the number of solved lines using an OLS regression (p-value= 0.733). We would not expect the degree of behindness aversion to have an effect in the *Low treatment*, but only in the *High treatment*. To account for this we add an interaction term between the treatment and the effort provision but the coefficient for alpha remains insignificant and so is the interaction term (p-value= 0.762). We obtain qualitatively similar results if we exclude the extremes on both sides ($\alpha = 0$ and $\alpha = 4.5$). Overall, the proportion of alphas is not far from what we expected and does not explain the lack of a treatment effect.

### 5.2.3 Leading Aversion

In our questionnaire, we have elicited a measurement of each subject's aversion to earning more than their peer. Initially, we did hypothesize that betas should not play a substantial role in the individual's decision of how much effort to provide. The betas we elicited in our experiment are on average 0.49 (where 6.98% have a beta of 0 and 8.72% have a beta of 1). If anything, we expected that workers with a higher degree of leading aversion should provide less effort. Surprisingly, this is not what we see in the data. First, if we run a simple OLS regression of the betas on the number of correctly solved lines we find a positive significant relationship (p-value= 0.009). However, if we exclude the extreme values ($\beta = 0$ and $\beta = 1$), this relationship vanishes. Comparing the distribution of betas to Beranek et al. (2015), we find that workers in their experiment are slightly more likely to report smaller values of beta but overall the distributions are very similar. In our experiment, 38.95% of the workers report a beta below 0.5, whereas in their experiment 54% do so (mean in our experiment = 0.49, mean in Beranek et al. (2015) = 0.48). Even though our measures for $\alpha$ and $\beta$ are not incentivized, we find overall very similar results using the same subject pool.

## 5.3 Further results

Here we report some additional analyses that go beyond our pre-registered main analysis. In our experiment, females tend to solve on average 20 lines more than males. This difference is large and statistically significant (Wilcoxon rank-sum test=0.029). We also find that more competitive workers solve more lines using an OLS regression (p-value= 0.005). It might be that more competitive people react more strongly to the treatment. To account for this we add an interaction term between the treatment and the degree of competitiveness but the interaction term is insignificant (p-value= 0.298). The gender composition of each pair does not seem to play a substantial role. There is no statistically significant difference in effort provision between a subject that is matched with someone of their gender and someone who is not (Wilcoxon rank-sum test=0.982). Similarly, the degree to which the workers feel connected does not seem to have a substantial effect. If we regress effort provision on the oneness, we find a weak and statistically insignificant relationship (p-value= 0.820).

Table 1: OLS Regressions: Effort provision as dependent variable

| | (1) | (2) | (3) |
|---|---|---|---|
| Treatment (1=high) | 1.774 | 1.741 | −0.170 |
| | (7.642) | (7.524) | (7.713) |
| Female | | 20.865*** | 22.464*** |
| | | (7.545) | (7.569) |
| Same gender | | −1.638 | −6.599 |
| | | (7.524) | (7.551) |
| Competitiveness | | | 9.100*** |
| | | | (2.703) |
| Oneness | | | −1.969 |
| | | | (4.189) |
| $\alpha_i$ | | | −1.581 |
| | | | (2.888) |
| $\beta_i$ | | | 29.246* |
| | | | (14.899) |
| Constant | 110.431*** | 99.843*** | 45.0376** |
| | (5.030) | (7.329) | (17.229) |
| Observations | 180 | 180 | 171 |

The table reports estimates from OLS regressions. The dependent variable is the number of solved lines in the real-effort task. No workers are excluded from the analysis in column (1) and (2). In column (3) workers with multiple switchpoints for our measures of $\alpha$ or $\beta$ are excluded. Standard errors are in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Our results are supported by the OLS regressions in Table 2. We regress the number of solved lines for each subject on a treatment dummy (see Table 2, column 1). The treatment difference is insignificant and stays insignificant when we control for observable characteristics (see Table 2, column 2) and the non-observable characteristics we elicited in the questionnaire (see Table 2, column 3). We control for the subject's gender, the gender composition of the group, the subject's degree of competitiveness, the perceived oneness of the group and our measures of behindness and leading aversion. The only highly significant control variables are gender and the level of competitiveness. This again suggests that females and more competitive people worked harder in our experiment.

We checked for gender differences in the reaction to the treatment manipulation. Men reacted positively to the treatment manipulation as they increased their average level of effort provision by 12.63 lines, women on the other hand decreased their average level of effort by 7.22 lines. However, the differences between the *High treatment* and the *Low treatment* are statistically insignificant for both subgroups (two-sided Wilcoxon rank-sum tests, p-value= 0.303 respectively 0.369).

# 6 Discussion & Recalibrated Experiment

## 6.1 Discussion

The workers in our experiment clearly solved more lines than expected given the findings of Gagnon et al. (2020). Whereas in Gagnon et al. (2020), workers solved on average 44 lines correctly, they solved more than 2.5 times as many in our experiment. There are different potential reasons why this may have happened. First, we used different subject pools. Whereas Gagnon et al. (2020) conducted their study on Prolific, we used the student subject pool of CeDEx. Another and potentially more important reason than the subject pool itself is that workers on Prolific can enter Prolific and search for experiments that are currently online. In contrast to that, we invited workers via Orsee and told them that the experiment could take up to two hours. It is likely that some workers on Prolific would need to leave the experiment at some point due to time restrictions whereas this is less likely for our experiment. Further, some workers on Prolific may have the possibility to leave the experiment and instead participate in another one which makes it more likely for them to quit if they expect higher earnings from switching.

The difference in effort levels between our study and the study of Gagnon et al. (2020) is problematic because we assume similar levels of effort for the calibration of our experiment. For this reason, we re-calibrated the working task by making the lines grow faster in the course of the experiment (now they increase by two characters instead of one character every five correctly solved lines). This way we are able able to observe a larger fraction of our workers that are below the high social reference point.

## 6.2 Recalibrated experiment

### 6.2.1 Main Analysis

We ran the identical experiment again with the exception that the lines increased now by two characters instead of one character for each five correctly solved lines. In total, the experiment consisted of 250 subjects of whom 125 subjects participated in the work task and 125 were passive peers. We collected 67 observations for the *High treatment* and 58 observations for the *Low treatment* respectively. The randomization to treatments lead to 63.8% females in the *High treatment* and 55.2% in the *Low treatment*. In both treatments, roughly 50% of the workers were paired with a peer of the same gender (50% in *Low* and 50.7% in *High*). All sessions were conducted online with students of the University of Nottingham in July 2021.

As in the original experiment, we do not find statistically significant differences between the treatments in terms of effort provision. In the *Low treatment*, the workers solved on average 57.65 lines correctly. In the *High treatment*, the workers solved on average 62.41 lines. Again, we use a one-sided Wilcoxon rank-sum test to test for the equality of effort provision between treatments yielding a p-value of 0.270. As in the original experiment, we do not find this effect even though workers in the *High treatment* are significantly less happy (two-sided Wilcoxon rank-sum test, p-value= 0.053). The treatment difference in terms of happiness is even stronger once we control for the individual performance in the working task in an OLS regression (p-value=0.019).

**Result 3:** The workers do not work significantly harder in the *High treatment* compared to the *Low treatment* in the recalibrated experiment.

Figure 5 shows a histogram of correctly solved lines for each treatment (*Low treatment* in the top panel, *High treatment* in the bottom panel). Again, one can see that the stopping decisions are dispersed over a wide range. Some workers stop directly or very early, however, this time no subject was able to solve all 180 lines. It is easy to see that the distribution of effort is shifted to the left compared to the original experiment. Again, the distributions of effort are very similarly dispersed between treatments (two-sided Kolmogorov–Smirnov test, p-value=0.506).

Our subsidiary hypothesis predicts a higher probability of stopping when the accumulated earnings equal the peer's earnings in the *High treatment*. We are interested to see whether there are systematic differences in terms of clustering

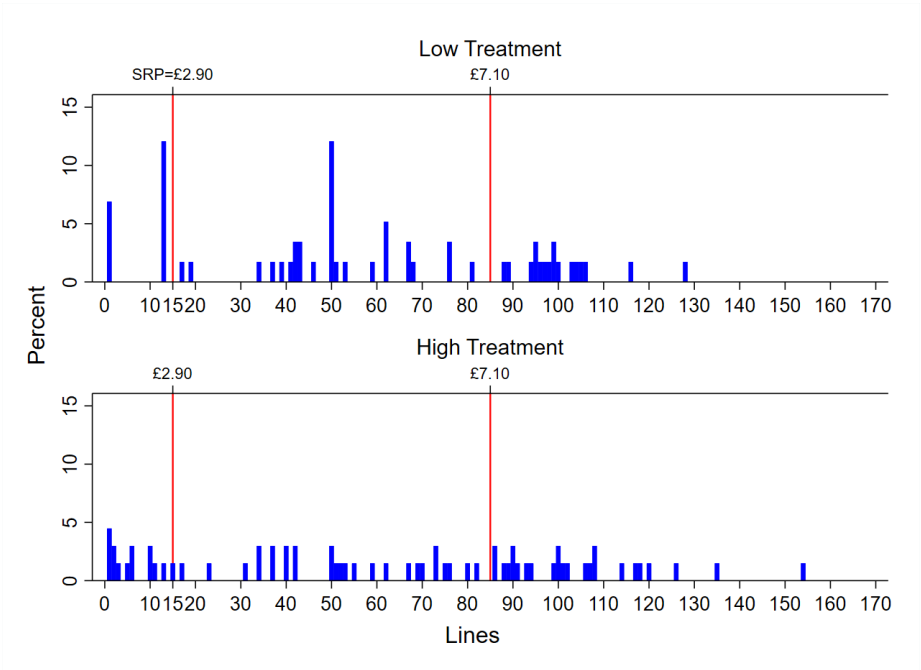



Figure 5: Histograms of correctly solved lines by treatment where the social reference points are indicated by the red lines in the re-calibrated experiment

of stopping decisions exactly at the peer's earnings: In neither treatment does a subject stop at exactly 85 lines, which corresponds to the income of the peer in the *High treatment*. If we compare the number of workers stopping in the range between 80 and 90 correctly solved lines, we find no significant differences between treatments (one-sided Fisher's exact test, p-value= 0.442). This also holds true for the ranges between 75 and 95 (one-sided Fisher's exact test, p-value= 0.172) and 70 and 100 correctly solved lines (one-sided Fisher's exact test, p-value= 0.509).

**Result 4:** The probability for the workers to stop at earnings = size of the high reference points is not higher in the *High treatment* compared to the *Low treatment* in the recalibrated experiment.

### 6.2.2 Further Results

Table 2: OLS Regressions: Effort provision as dependent variable

| | (1) | (2) | (3) |
|---|---|---|---|
| Treatment (1=high) | 4.763 | 5.990 | 3.266 |
| | (6.774) | (6.443) | (6.851) |
| Female | | 15.797** | 14.947** |
| | | (6.755) | (7.092) |
| Same gender | | 16.955** | 16.983** |
| | | (6.614) | (7.067) |
| Competitiveness | | | 4.777** |
| | | | (2.270) |
| Oneness | | | 0.298 |
| | | | (3.818) |
| $\alpha_i$ | | | −0.240 |
| | | | (2.713) |
| $\beta_i$ | | | 5.814 |
| | | | (13.374) |
| Constant | 57.655*** | 39.100*** | 16.463 |
| | (4.960) | (6.662) | (15.500) |
| Observations | 125 | 125 | 116 |

The table reports estimates from OLS regressions. The dependent variable is the number of solved lines in the real-effort task. No workers are excluded from the analysis in column (1) and (2). In column (3) workers with multiple switchpoints for our measures of $\alpha$ or $\beta$ are excluded. Standard errors are in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

As in our original experiment, subjects remembered the earnings of their peer fairly well - 80% of the workers remember the exact size of their peer's earnings and 98.4% if we allow for a deviation of ± 1. Again, we find no effect of the degree of behindness aversion on the level of effort provision using a simple OLS regression (p-value= 0.779) and this lack of treatment effect continues once we add an interaction term (p-value= 0.423). In contrast to the original experiment, we also find no effect of individual betas on the number of correctly solved lines (p-value= 0.376). Again, females work significantly harder than males using a

two-sided Wilcoxon rank-sum test (p-value= 0.005). Similarly to our original experiment we find that more competitive workers supply more effort (p-value= 0.050) and the degree of oneness does not seem to play a substantial role in determining effort provision using OLS regressions (p-value= 0.634). In contrast to our original experiment, workers who have been matched with a person of the same gender work significantly harder (Wilcoxon rank-sum test, p-value= 0.018). These results are confirmed by the regression in Table 3.

Again we check for heterogeneous treatment effects for men and women. In contrast to our original experiment, women solved on average 11.64 lines more when being allocated to the *High treatment*. As in our original experiment, this difference is not statistically significant (two-sided Wilcoxon rank-sum test, p-value= 0.251). Also in contrast to our original experiment, men worked slightly less in the *High treatment* compared to the *Low treatment* and again this difference is not statistically significant (two-sided Wilcoxon rank-sum test, p-value= 0.592).

# 7 Conclusion

Using a simple laboratory experiment, we test whether social reference points influence real effort provision in a working task. Our experiment is based on a simple model of social reference points, following Fehr and Schmidt (1999). We design an experimental paradigm that allows us to measure the level of effort provision precisely while exogenously manipulating a reference point. This allows us to identify the effect of a social reference point by comparing the level of effort provision between treatments.

In our experiment, only two subjects participate in each session. Each subject is assigned one of two roles worker or peer, by a publicly observed coin toss. The peer receives a fixed payment, whereas the worker participates in a real effort task. The fixed payment to the peer is varied between the two treatments. This allows us to trace any difference in the average level of effort provided by the workers across treatments back to social reference points.

We find that the workers in our experiment care about the earnings of their respective peers, i.e. they remember the earnings of their peers and are less happy with their earnings when their peer earns relatively more. Despite this, we cannot observe a change in working behavior. The workers in our experiment do not increase their level of effort when faced with relatively high peer earnings

compared to low reference earnings. This suggests that people care about income differences but this does not necessarily translate to a change in behavior in incentivized environments.

To test the robustness of our findings we conducted a follow-up experiment with a modified and more difficult working task. The results are in line with our original experiment. Again, workers were significantly less happy when their peer earned relatively more. As in our original experiment, the difference in satisfaction did not lead to a significant treatment effect in terms of real effort provision in the work task.

We now turn to the issue of why we do not observe a treatment effect. Festinger (1954) discusses what makes a reference point salient. He argues that people who are most like you are the most influential point of comparison. While our participants are very similar in many ways, they differ in how they can earn their income. Whereas workers earn their income by providing effort, the income of peers is determined by luck. One possible explanation for why we do not find differences in effort provision between the treatments could be that workers are less likely to use the income of a peer as a relevant point of comparison if it does not result from the same work task. However, it remains puzzling that we do find that workers care about the income of their peers but it did not translate to changes in economic behavior.

It is interesting to compare our results with the literature on wage inequalities that suggests that others' earnings might modify behavior. For instance, Bracha et al. (2015) and Gagnon et al. (2020) find that workers work less when they earn less than others for the same work task. However, this effect can not be attributed to relative income comparisons but must be interpreted as a net effect of wage discrimination. It is possible that there is a positive effect of income targeting on effort provision hidden under a stronger negative effect of discrimination. Our experiment shuts down several channels that might influence effort provision and cleanly tests the effect of relative income comparisons on effort provision.

One possible interpretation of our results is that fixed peer outcomes do not serve as reference point. However, there is some evidence that salient peer outcomes in the way we implement them in our experiment can serve as reference points and modify behavior. Schwerter (2023) finds that people are more willing to take risks when they face higher social reference points using a similar design. The experiments of Schwerter (2023) and us differ in that workers in his experiment make only one payoff-relevant decision about a lottery, while in

our experiment they work for up to 90 minutes. Social reference points as used in our experiments may have short rather than long-term effects on economic behavior. A possible explanation would be that social reference points might be a transient thing where the duration and size of the effect depend on how strong the point of comparison is. However, the workers in our experiment remembered the social reference point by the end of the experiment and cared about it at least on a psychological level, so it is hard to view the effects of the social reference point in our experiment as transient.

For future research, it would be interesting to investigate in more detail where and why relative incomes serve as social reference points, and how this affects labour supply behavior.

# References


Abeler, J., A. Falk, L. Goette, and D. Huffman (2011). Reference points and effort provision. *American Economic Review 101*(2), 470–92.

Akerlof, G. A. and J. L. Yellen (1990). The fair wage-effort hypothesis and unemployment. *The Quarterly Journal of Economics 105*(2), 255–283.

Austin, W., N. C. McGinn, and C. Susmilch (1980). Internal standards revisited: Effects of social comparisons and expectancies on judgments of fairness and satisfaction. *Journal of Experimental Social Psychology 16*(5), 426–441.

Bandura, A. (1977). Self-efficacy: toward a unifying theory of behavioral change. *Psychological Review 84*(2), 191.

Beranek, B., R. Cubitt, and S. Gächter (2015). Stated and revealed inequality aversion in three subject pools. *Journal of the Economic Science Association 1*(1), 43–58.

Blanco, M., D. Engelmann, and H. T. Normann (2011). A within-subject analysis of other-regarding preferences. *Games and Economic Behavior 72*(2), 321–338.

Bracha, A., U. Gneezy, and G. Loewenstein (2015). Relative pay and labor supply. *Journal of Labor Economics 33*(2), 297–315.

Card, D., A. Mas, E. Moretti, and E. Saez (2012). Inequality at work: The effect of peer salaries on job satisfaction. *American Economic Review 102*(6), 2981–3003.

Clark, A. E. and A. J. Oswald (1996). Satisfaction and comparison income. *Journal of Public Economics 61*(3), 359–381.

Cohn, A., E. Fehr, B. Herrmann, and F. Schneider (2014). Social comparison and effort provision: Evidence from a field experiment. *Journal of the European Economic Association 12*(4), 877–898.

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

Fehr, E. and S. Gächter (2000). Fairness and retaliation: The economics of reciprocity. *Journal of Economic Perspectives 14*(3), 159–181.

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

Ferreri-Carbonell, A. (2005). Income and well-being: an empirical analysis of the comparison income effect. *Journal of Public Economics 89*(5-6), 997–1019.

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

Gächter, S., L. Huang, and M. Sefton (2018). Disappointment aversion and social comparisons in a real-effort competition. *Economic Inquiry 56*(3), 1512–1525.

Gächter, S., C. Starmer, and F. Tufano (2015). Measuring the closeness of relationships: a comprehensive evaluation of the'inclusion of the other in the self'scale. *PLOS ONE 10*(6), e0129478.

Gächter, S. and C. Thöni (2010). Social comparison and performance: Experimental evidence on the fair wage–effort hypothesis. *Journal of Economic Behavior & Organization 76*(3), 531–543.

Gagnon, N., K. Bosmans, and A. Riedl (2020). The effect of unfair chances and gender discrimination on labor supply. *Available at SSRN 3519540*.

Genicot, G. and D. Ray (2017). Aspirations and inequality. *Econometrica 85*(2), 489–519.

Gill, D. and V. Prowse (2012). A structural analysis of disappointment aversion in a real effort competition. *American Economic Review 102*(1), 469–503.

Greiner, B. (2015). Subject pool recruitment procedures: organizing experiments with orsee. *Journal of the Economic Science Association 1*(1), 114–125.

Kőszegi, B. and M. Rabin (2006). A model of reference-dependent preferences. *The Quarterly Journal of Economics 121*(4), 1133–1165.

Luttmer, E. F. (2005). Neighbors as negatives: Relative earnings and well-being. *The Quarterly Journal of Economics 120*(3), 963–1002.

Marmot, M. (2005). *Status syndrome: How your social standing directly affects your health*. A&C Black.

Mas, A. and E. Moretti (2009). Peers at work. *American Economic Review 99*(1), 112–45.

Perez-Truglia, R. (2020). The effects of income transparency on well-being: Evidence from a natural experiment. *American Economic Review 110*(4), 1019–54.

Sacerdote, B. (2001). Peer effects with random assignment: Results for dartmouth roommates. *The Quarterly Journal of Economics 116*(2), 681–704.

Schwerter, F. (2023). Social reference points and risk taking. *Management Science*.

Tversky, A. and D. Kahneman (1979). Prospect theory: An analysis of decision under risk. *Econometrica 47*(2), 263–291.

# Appendices

## A Power Analysis

We conducted a power analysis based on the following structural model. We assume subject $i$ gets utility from:

$$U_i = \begin{cases} R + we_i - \alpha_i(F - R - we_i) - \frac{\lambda_i e_i^2}{2} & \text{if } R + we_i < F \\ R + we_i - \frac{\lambda_i e_i^2}{2} & \text{if } R + we_i \geq F \end{cases} \tag{5}$$

where $R$ is the show-up fee and $w$ the piece-rate wage. With this specification optimal effort is,

$$e_i^* = \begin{cases} \frac{(1+\alpha_i)w}{\lambda_i} & \text{if } R + \frac{(1+\alpha_i)w^2}{\lambda_i} \leq F \\ \frac{F-R}{w} & \text{if } R + \frac{w^2}{\lambda_i} < F < R + \frac{(1+\alpha_i)w^2}{\lambda_i} \\ \frac{w}{\lambda_i} & \text{if } \frac{w^2}{\lambda_i} + R \geq F \end{cases} \tag{6}$$

In the absence of reference points, or if $a_i = 0$ optimal effort is simply $e_i = \frac{w}{\lambda_i}$. However, if $a_i > 0$ the worker gets disutility from earning less than the paired subject, and this results in higher effort. Moreover, the model predicts higher effort in the high reference point treatment.

We calibrate the model as follows. First, using the results from Gagnon et al. (2020) we assume $\frac{w}{\lambda_i} \sim \mathcal{N}(44.03, 28.93)$. We use a very similar real effort task as them and use the same piece-rate parameters to their treatment which results in average effort of 44.03 lines with a standard deviation of 28.93 lines (see Table 4 of Gagnon et al. (2020)).[4] If anything, we expect that the workers solve slightly more lines in our experiment. To account for this we will present power estimates for the results from Gagnon et al. (2020) and for the case that the average effort and standard deviation increase by 10% compared to Gagnon et al. (2020), i.e. $\frac{w}{\lambda_i} \sim \mathcal{N}(48.43, 31.82)$.

Our task parameters are $w = 0.06$ per line and $R = 2$. Next, we take a slightly conservative approach based on Fehr and Schmidt (1999) and assume that one-third of the population has $a_i = 0$, one-third has $a_i = 0.5$ and one-third

[4] Gagnon et al. (2020) impose an upper bound of 85 correctly solved lines. Unlike Gagnon et al (2020), we increase the upper bound to 170 lines. Also, we increased the amount of time for the working task from 70 to 90 minutes.

has $a_i = 1$.[5] Further we assume all betas to be zero.

To estimate the power of the test we conducted a Monte Carlo simulation with 10000 replications for varying sample sizes ($n = 60, 90, 120, 150, 180$) with the reference earnings (($F_H$,$F_L$)) = (7.10,2.90). The sample size refers to the total number of workers completing the task in both treatments. We chose the reference points so that reference earnings are exactly attainable given our task parameters (i.e. $(F-2)/0.06$ is an integer) and workers taking part in the experiment but leaving with a fixed sum receive a minimum payment and expected payment per hour consistent with our lab standards.[6]

In each replication, we drew n observations, where each observation corresponds to an experimental subject and each subject is assigned $\lambda_i$ and $\alpha_i$ parameters by independent draws from the distributions described above. Each observation is then assigned $F = F_H$ or $F = F_L$ with equal probability to form two samples corresponding to workers assigned to complete the task with reference earnings $F_H$ or $F_L$ respectively. For each observation, effort is calculated using the optimal effort function.

To test our main hypothesis, we then conduct a one-sided Wilcoxon Mann-Whitney rank-sum test at the 5% significance level. The proportions of rejections in 10000 replications are our estimates of power. Table A1 reports estimated power for $w/\lambda_i \sim \mathcal{N}(44.03, 28.93)$ as in Gagnon et al. (2020) and $w/\lambda_i \sim \mathcal{N}(48.43, 31.82)$.

| n | $w/\lambda_i \sim \mathcal{N}(44.03, 28.93)$ | $w/\lambda_i \sim \mathcal{N}(48.43, 31.82)$ |
|---|---|---|
| 60 | 0.5 | 0.44 |
| 90 | 0.64 | 0.58 |
| 120 | 0.75 | 0.68 |
| 150 | 0.84 | 0.77 |
| 180 | 0.88 | 0.83 |

Table A1: Estimated Power for Main Hypothesis

Our subsidiary hypothesis is that more workers match the high reference earnings in the high reference point treatment compared to the low reference

[5] Fehr and Schmidt (1999) initially assume that 0.3 of the population has $a_i = 0$, 0.3 has $a_i = 0.5$ and 0.3 has $a_i = 1$ and 0.1 of the population has $a_i = 4$.
[6] We expect an average hourly wage of £8.10 for participating in the working task.

point treatment. For our subsidiary hypothesis, we conduct a one-sided Fisher's exact test at the 5% significance level. Again, the proportion of rejections in 10000 replications are our estimates of power. Table A2 reports the estimated power for $w/\lambda_i \sim \mathcal{N}(44.03, 28.93)$ and $w/\lambda_i \sim \mathcal{N}(48.43, 31.82)$. There is another subsidiary hypothesis that more workers match the earnings of the social reference point in the low reference point treatment. However, given our parameters, it is unlikely that this test will be rejected, e.g. for $n = 180$ the probability of rejection is less than 20%.

| n | $w/\lambda_i \sim \mathcal{N}(44.03, 28.93)$ | $w/\lambda_i \sim \mathcal{N}(48.43, 31.82)$ |
|---|---|---|
| 60 | 0.86 | 0.88 |
| 90 | 0.98 | 0.98 |
| 120 | 1 | 1 |
| 150 | 1 | 1 |
| 180 | 1 | 1 |

Table A2: Estimated Power for Subsidiary Hypothesis

Based on this analysis we concluded that the proposed test procedure has adequate power ($> 80\%$) with 180 workers completing the task. Note that this implies recruiting a total of 360 subjects (recall there is a paired subject for each subject completing the task).

## B Recruitment message

Dear fname,

You are registered with CeDEx to participate in experiments. We would like to invite you to take part in our upcoming experiment. The experiment will take place online and can take from 5 to 100 minutes. All payments will be done with PayPal.

For this experiment, you will need Microsoft Teams and your camera on for the first 10 minutes of the experiment. We will videocall you at the time you have registered for.

You will need to use a computer for this experiment. Please, do not try to use your mobile phones or tablets instead.

We are planning to run many sessions this week and more sessions might be added.

If you would like to participate, please click on the link below to sign up for the session of your choice. Please note: people that sign up for a session and do not turn up cause us problems. Please sign up only if you are sure you can attend it and if you sign up please do attend. We operate a policy of removing participants from our database who sign up but then fail to turn up.

#link#

(If the link does not work, copy it and paste it into the address field of your internet browser.)

Best regards,

CeDEx Team

## C Welcome Speech

At the beginning of each experimental session, the following welcome speech will be given by the experimenter:

*Welcome to our experiment today! The experiment consists of a simple work task! However, only one of you will participate in this task - Player A. The second person, Player B, will receive a fixed amount of money and will not participate in the work task. Both the amount of money and which one of you will participate in the work task will depend on luck! For this purpose, I will toss two coins. Both coin tosses will be visible for you. The first coin determines the amount of money, Player B will receive. If the coin shows £7.10, Player B will receive £7.10. If the coin shows £2.90, Player B will receive £2.90. Then I will toss another coin. If the coin shows green, [first name of second participant] will be Player A and will take part in the work task. If the coin shows blue, [first name of second participant] will be Player A and will take part in the work task. After that, Player A will receive a link that leads him to the instructions of the work task and both of you can leave the videocall.*

# D Instructions for Player A

**Instructions**

Welcome to this economic experiment! In this experiment you earn money by participating in a working task. These instructions will describe the task and how you earn money, so please read the instructions carefully.

**The working task**

The working task consists of entering lines of text on your computer. You will see one line at a time. Each time you see a new line, you can decide whether to type this line or leave the experiment.

If you decide to type the line you must type it correctly before going on to the next page to see the next line. In case you make a mistake when entering the line, the software will tell you so and you will have to type it again.

If you decide to leave the experiment this will end the working task, you will be asked to fill out a short questionnaire (which will take about five minutes), and you will be informed of your final payment. Note that if you decide to leave, you will not be able to start working again. That is, once you leave the working task you cannot go back.

Each time you are presented a new line you will have the option of typing it or leaving the experiment. The length of the lines will increase as you complete more lines. After each five lines are correctly entered, the length of a line increases by two characters.

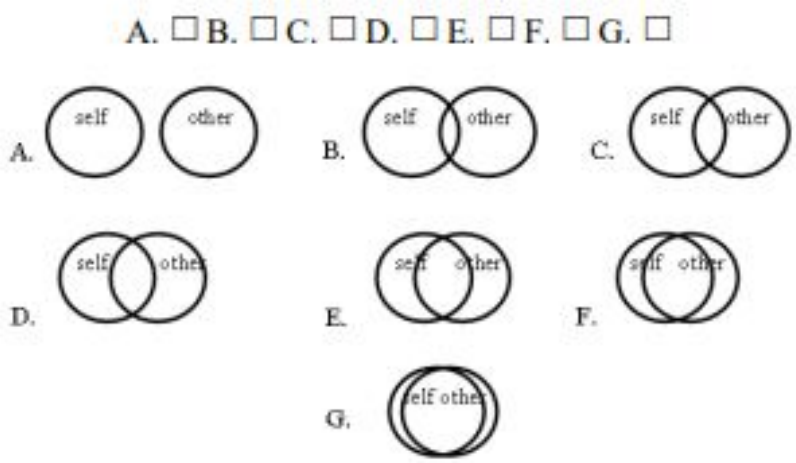
You have up to 90 minutes to work on the task. However, you can finish earlier if you want by choosing to leave the experiment when you are presented with a new line.

**How you earn money**

When you leave the experiment according to the described procedure you will be informed of your payment. You will receive a payment of £0.06 for each line you entered correctly. In addition, you will receive a fixed amount of £2.00, irrespective of the number of lines entered.

## E Questionnaire

1. What is your gender? [male, female, neutral]
2. What is the gender of player B? [male, female, neutral, I don't know]
3. How much money did Player B receive?
4. Please, look at the circles diagram provided on your desk. Then, consider which of these pairs of circles best represents your connection with this person before this experiment. By selecting the appropriate letter below, please indicate to what extent you and Player B were connected.




5. How competitive do you consider yourself to be?" [1 (not competitive at all) -7 (very competitive)]
6. How happy are you with your earnings in today's experiment? [1 (not happy at all) -7 (very happy)]

Imagine the following hypothetical scenario: You are asked to choose between two possible allocations of money between you and another person in eleven different decision problems as presented below. Please indicate for each row which decision you prefer.

| | What do you prefer? | | |
|---|---|---|---|
| You receive £0, the other person receives £0 | ○ | ○ | You receive £0, the other person receives £20 |
| You receive £0, the other person receives £0 | ○ | ○ | You receive £1, the other person receives £19 |
| You receive £0, the other person receives £0 | ○ | ○ | You receive £2, the other person receives £18 |
| You receive £0, the other person receives £0 | ○ | ○ | You receive £3, the other person receives £17 |
| You receive £0, the other person receives £0 | ○ | ○ | You receive £4, the other person receives £16 |
| You receive £0, the other person receives £0 | ○ | ○ | You receive £5, the other person receives £15 |
| You receive £0, the other person receives £0 | ○ | ○ | You receive £6, the other person receives £14 |
| You receive £0, the other person receives £0 | ○ | ○ | You receive £7, the other person receives £13 |
| You receive £0, the other person receives £0 | ○ | ○ | You receive £8, the other person receives £12 |
| You receive £0, the other person receives £0 | ○ | ○ | You receive £9, the other person receives £11 |
| You receive £0, the other person receives £0 | ○ | ○ | You receive £10, the other person receives £10 |

---

Imagine the following hypothetical scenario: You are asked to choose between two possible allocations of money between you and another person in eleven different decision problems as presented below. Please indicate for each row which decision you prefer.

| | What do you prefer? | | |
|---|---|---|---|
| You receive £10, the other person receives £0 | ○ | ○ | You receive £0, the other person receives £0 |
| You receive £10, the other person receives £0 | ○ | ○ | You receive £1, the other person receives £1 |
| You receive £10, the other person receives £0 | ○ | ○ | You receive £2, the other person receives £2 |
| You receive £10, the other person receives £0 | ○ | ○ | You receive £3, the other person receives £3 |
| You receive £10, the other person receives £0 | ○ | ○ | You receive £4, the other person receives £4 |
| You receive £10, the other person receives £0 | ○ | ○ | You receive £5, the other person receives £5 |
| You receive £10, the other person receives £0 | ○ | ○ | You receive £6, the other person receives £6 |
| You receive £10, the other person receives £0 | ○ | ○ | You receive £7, the other person receives £7 |
| You receive £10, the other person receives £0 | ○ | ○ | You receive £8, the other person receives £8 |
| You receive £10, the other person receives £0 | ○ | ○ | You receive £9, the other person receives £9 |
| You receive £10, the other person receives £0 | ○ | ○ | You receive £10, the other person receives £10 |