This is a previous version of the article published in Management Science. 2022, 68(10): 7518-7537. https://doi.org/10.1287/mnsc.2021.4280

Loss-framed incentives and employee (mis-)behavior*

Eszter Czibor[1], Danny Hsu[2], David Jimenez-Gomez[3], Susanne Neckermann[4] and Burcu Subasi[5]

[1]Innovation Growth Lab, Nesta
[2]Erasmus School of Economics
[3]FAE, University of Alicante
[4]University of Chicago
[5]University of Groningen

November 13, 2021

Abstract

This paper explores how loss-framed incentives affect behavior in a multitasking environment in which participants have more than one way of recovering (expected) losses. In a real-effort laboratory experiment, we offer participants task incentives that are either framed as a reward (gain) or as a penalty (loss). We study their responses along three dimensions: performance in the incentivized task, theft, and voluntary provision of help. We find that framing incentives as a penalty rather than as a reward does not significantly improve task performance, but increases theft and leads to a small and insignificant reduction in the share of participants willing to help the experimenter. Secondary analyses based on our theoretical framework help us pin down the mechanism at play and suggest that loss aversion drives participants' response. Our findings have important implications for incentive design in practice.

JEL-codes: D91, M52, C92
Keywords: loss-framed incentives; multitasking; incentive design; stealing

*Acknowledgments: We would like to thank seminar participants at the University of Chicago, the Ludwig Maximilian University of Munich, the Rady School of Management, and the University of Regensburg for their valuable insights. We are particularly grateful for advice from Pedro Albarran, Leonardo Bursztyn, Uri Gneezy, Alex Imas, Juanna Joensen, Steve Levitt, John List, Gert Poenitzsch, Sally Sadoff, and Timo Vogelsang. David Jimenez-Gomez is grateful for the funding from the Ministry of Science and Innovation and the European Regional Development Fund (FEDER), through the PID2019-107081GB-I00 project; from the Ministry of Science and Innovation, through the project "Nudging applied to the improvement of regulation" (RED2018-102761-T); and from the Valencian Community, through the Prometeo program PROMETEO/2021/073.

1 Introduction

Loss-framed incentives have received a lot of attention in behavioral economics. In contrast to traditional gain-framed incentive schemes where employees receive a reward upon achieving a pre-specified target, loss-framed incentive schemes entail an upfront payment to the employees, which they lose if they fail to reach the target.[1] Several papers have found that loss-framed incentives lead to greater effort provision than their gain-framed equivalents (e.g. Armantier and Boly, 2015; Bulte et al., 2019; Fryer et al., 2012; Hannan et al., 2005; Hossain and List, 2012; Levitt et al., 2016). In this literature, employees typically have only one way of reducing their expected losses, by increasing their actual or reported effort on the task.

We contribute to this literature by exploring how loss-framed incentives affect behavior in a multitasking environment where employees have more than one strategy available to reduce their expected losses. In particular, we study a setting where employees can respond to incentives by exerting effort on the incentivized task, stealing from their employer and/or withholding voluntary effort on a different, non-incentivized task that benefits their employer. Theft and helping represent employee behaviors that are typically not directly incentivized but that have a crucial impact on an organization's success. Theft, for example, poses enormous costs to businesses worldwide: Hermann and Mußhoff (2019) report that businesses suffer about $48 billion of retail loss annually due to employee theft. Voluntary helping behavior, by comparison, entails activities that go above and beyond that which is formally required by employees' job descriptions and are considered essential for the success of organizations as a whole (Bradler and Neckermann, 2016; Harbring and Irlenbusch, 2011; Neckermann et al., 2014).[2]

We study the impact of loss-framed incentives on these behaviors in a laboratory experiment where we randomize the framing of the monetary incentives in a real-effort task. Participants assigned to the *Reward* treatment earn money for every correct answer given on a matrix task, while participants in the *Penalty* treatment start with an endowment and lose money every time they make a mistake. The two payment schemes are payoff equivalent. We treat task performance as an indicator for effort provision. Upon completion of the task, participants are informed that they can help the experimenter with a survey (our proxy for voluntary helping), and are required to fill out an obligatory questionnaire (which measures participants' satisfaction with various aspects of the experiment). We measure stealing using a novel experimental approach: we seat participants in completely isolated cubicles to minimize any perception of scrutiny, place a box full of various office supply items (including pens and pencils to be used to fill out the survey) on each desk, and count after the experiment whether any items are missing. By comparing task performance, theft and survey completion between the two treatments, we can study how loss-framed incentives affect behavior in a multidimensional setting

[1] Gain-framed incentives provide a useful benchmark to loss-framed incentives as long as the two incentive schemes are payoff equivalent and the only difference between the two treatments is the framing as a reward or penalty.

[2] Examples include organizing team events, providing constructive feedback to a colleague, or substituting for sick employees – activities collectively referred to as "organizational citizenship behavior" (Podsakoff et al., 2000).

and investigate the underlying mechanisms.

We explore these mechanisms using our theoretical framework based on the multitasking model of Pierce et al. (2020). Our framework incorporates two mechanisms that may govern employees' response to loss-framed incentives: loss aversion and behavioral spillovers (Grolleau et al., 2016). The *loss aversion* channel entails two key components (Thaler and Johnson, 1990): first, losses loom larger than gains; second, participants in the *Penalty* treatment view the upfront payment received as their reference point, and thus have a higher reference point than participants in the *Reward* treatment who receive no initial payment (see Section 3.2 for details on the experiment design regarding the reference point). In single effort settings, this channel implies that employees work harder in an attempt to avoid or reduce their losses than they do in order to achieve gains. In multidimensional settings, however, employees may have access to more than one strategy to avoid losses, not all of which involve greater effort provision (Pierce et al., 2020). Loss-framed incentives might also affect non-incentivized behaviors by reducing the moral cost of selfish and unethical behavior – a phenomenon we refer to as the *behavioral spillovers* channel. There are various reasons why we might expect loss-framed incentives to cause behavioral spillovers. First, they may provide "moral wiggle room" to justify unethical behavior (Dana et al., 2007; Mazar et al., 2008; Rabin, 1994).[3] Second, the incentive scheme could affect the prevailing norms, and, hence, participants' value orientation (Bowles, 1998; Goette et al., 2012).[4] Third, loss-framed incentives may cause participants to feel negative emotions such as anger and frustration, which, in turn, affect subsequent behavior (Gneezy and Imas, 2014; Koszegi, 2006; Loewenstein, 2000). If participants blame the experimenter for the negative emotions they experience, loss-framed incentives may reduce their altruism toward the experimenter and lead to an increase in theft or a decrease in the voluntary provision of help (Dur, 2009; Fehr and Schmidt, 2006).[5]

While both the loss aversion and behavioral spillover channels predict a higher prevalence of theft in the *Penalty* than in the *Reward* treatment, they differ in the explanation they provide. According to the loss aversion channel, participants work harder and/or steal in order to eliminate their losses, suggesting a bunching in combined income from task earnings and theft just above their target earnings in the *Penalty* treatment, while the behavioral spillovers channel predicts a universal shift in the distribution of theft to the right. Moreover, both channels predict a reduction in voluntary helping in response to loss-framed incentives. In the case of the loss aversion channel this prediction arises from increased incentives for income-

[3] Receiving an upfront endowment may increase participants' sense of entitlement and deservingness, which in turn may justify cheating (Schindler and Pfattheicher, 2017; Shalvi et al., 2011) or theft (Cameron and Miller, 2009; Gravert, 2013; Schurr and Ritov, 2016).

[4] Gneezy et al. (2011) provide numerous examples in support of the claim that the framing of a decision situation critically influences pro-social behavior. Buser and Dreber (2016) show that a competitive prime alone – without actual competitive incentives – may reduce cooperation. Such a change in norms may even make unethical behavior "more acceptable not only by the individual but also by third parties"(Grolleau et al., 2016, p.3435).

[5] Relatedly, Breza et al. (2018) find that pay inequality only hurts output, attendance, and group cohesion when the source of this inequality – differences in worker output – is not readily observable, leading to perceived fairness violations. (Ockenfels et al., 2015) find that when employees become dissatisfied when their bonus payment falls below a natural reference standard for a fair bonus.

generating activities (i.e. task effort and theft) that decrease the relative incentives for survey completion. According to the behavioral spillovers channel, the reduction in helping is a direct consequence of the assumed reduction in the utility from completing the survey.

Our empirical results show that framing incentives as losses rather than as gains has a negligible effect on participants' performance in the incentivized task, but a large impact on the prevalence of theft in our experiment. While the difference in mean task scores between the two treatments is small and not statistically significant, the share of participants who stole something is more than twice as high in the *Penalty* than in the *Reward* treatment. We observe a moderate increase in the average size of theft: the mean value of items stolen (among all participants, including those who did not steal) is 44% higher in the *Penalty* than in the *Reward* treatment, although the difference between the treatments is not statistically significant.[6] Participants in the *Penalty* treatment are somewhat (3.6 percentage points) less inclined to complete the voluntary survey, but the estimated difference is not statistically significant. Studying the distribution of participants' combined income from task earnings and theft, we find evidence suggestive of bunching just above participants' target earnings in the *Penalty* treatment. We find no meaningful difference between the two treatments in terms of participants' self-reported level of satisfaction with the experiment. These results are largely consistent with the predictions of the loss aversion channel outlined above, and they do not offer convincing evidence in support of the behavioral spillover explanation.

Our paper makes four distinct contributions to the literature. First, it improves our understanding of how loss-framed incentives affect employee behavior in complex environments. Several papers measure the impact of loss-framed incentives on employees' effort and task performance in various settings (e.g. Armantier and Boly, 2015; Brooks et al., 2012; Bulte et al., 2019; De Quidt et al., 2017; DellaVigna and Pope, 2017; Fryer et al., 2012; Hannan et al., 2005; Hossain and List, 2012; Levitt et al., 2016). A small but growing literature extends the analysis to dishonest means of increasing one's earnings and shows that loss-framed incentives tend to increase unethical behavior (Cameron and Miller, 2009; Grolleau et al., 2016; Kern and Chugh, 2009; Pettit et al., 2016; Schindler and Pfattheicher, 2017; Shalvi, 2012). Our study combines these two strands of the literature: we ask whether loss-framed incentives induce higher effort provision in a multitasking environment where employees have access to both honest *and* dishonest ways of eliminating their losses.[7]

[6] The average task scores are 85.7 and 86.2 out of 100 in the *Reward* and *Penalty* treatment, respectively; a difference that corresponds to less than 1% of the mean in the *Reward* treatment, or approx. 5% of the pooled standard deviation. In the *Reward* treatment, only 11.3% (18 out of the 159 participants) steal anything, while the corresponding share is 23.6% (38 out of 161) in the *Penalty* treatment (p-value of two-sample test of proportions: 0.004). In the *Reward* treatment, participants steal items worth 0.47€ on average, while the mean value of items stolen is 0.67€ in the *Penalty* treatment (both means are calculated over all participants, including those who stole nothing). The p-value from a t-test comparing the mean value stolen between treatments is 0.336.

[7] There are a few existing studies on unethical behavior and loss-framed incentives that allow participants to exert more effort on the task as well as to cheat or to steal. In these studies, however, researchers can only observe how loss framing affects participants' *reported* performance (Grolleau et al., 2016) or the amount of money they take (Cameron and Miller, 2009), but they cannot decompose participants' response into a change in *actual* performance vs. a change in dishonest/immoral behavior. In contrast, our design helps us to disentangle increased effort provision from more dishonesty not only on the group but also on the individual level, and thus

Second, our study contributes to the literature on incentives in multitasking settings (Holmstrom and Milgrom, 1991). Our results echo the findings of Pierce et al. (2020) who show that loss framing may exacerbate the incentives for an "undesirable allocation of effort across dimensions." In particular, Pierce et al. (2020) find that car dealers in a field experiment respond to loss-framed incentives by allocating their effort across multiple dimensions in a way that helps them mitigate their exposure to losses but reduces overall revenue. While their study provides compelling field evidence that loss framing might induce cross-dimension gaming that leads to lower overall revenue, our experiment shows that employees may respond to loss-framed incentives by increasing non-incentivized behavior that is directly harmful for the employer.[8]

Third, our paper improves our knowledge of the factors driving theft. Despite theft being a large and costly challenge for organizations, we know of only a handful of studies that consider it in controlled experiments. Gravert (2013) shows that people steal more when their payoffs are based on performance rather than on luck, while Belot and Schröder (2016) find that monitoring productivity and penalizing mistakes does not increase theft.[9] Our results confirm that theft is not determined by individuals' moral cost alone but is responsive to contextual factors (Pierce et al., 2015). In our case, a simple change in the framing of the incentive led to a large increase in the share of people who stole.

Finally, we advance experimental methodology by introducing a novel paradigm to measure theft in a laboratory setting. In particular, we place a large box of office supplies on the desk in each participants' cubicle, and record the number and value of any items missing from the container after the experiment. Existing research operationalizes theft either as taking more money than deserved (Cameron and Miller, 2009; Gravert, 2013; Hermann and Mußhoff, 2019) or using a task that involves sending participants home with boxes of euro coins (whose contents they need to identify) and subsequently assessing whether coins are missing from the boxes after they are returned (Belot and Schröder, 2013, 2016). We believe that our method complements previous approaches in several ways. First, due to its inconspicuousness, it may reduce experimenter demand effects. Second, it may allow for what Hsee et al. (2003) and Mazar et al. (2008) call "malleable categorization of behavior," the idea being that it might be easier

allows us to explore the underlying drivers of participants' behavior.

[8] Our experiment is also related to studies of the quantity-quality trade-off in multitasking environments. By allowing our participants to allocate their effort between an incentivized and a non-incentivized task (matrix task vs. voluntary survey), our design captures the idea that while some dimensions of effort provision lead to easily observable outcomes and can be directly incentivized (similar to the quantity dimension), others are hard to observe or contract (similar to the quality dimension). Hossain and List (2012) find no effect of loss-framed incentives on the quality of output in a field experiment among factory workers, while Rubin et al. (2018) find evidence for a quantity-quality trade-off in the lab. Studying financial incentives in the context of non-routine, analytical team tasks, Englmaier et al. (2018) document a positive impact on performance and a negative impact on the willingness to explore new solutions (an aspect of quality) – irrespective of the incentives being framed as gains or losses.

[9] Belot and Schröder (2013) study the relationship between competitive incentives and stealing, but observe hardly any theft at all. Cameron and Miller (2009) are interested in whether people steal more to avoid losses than to achieve gains, measuring stealing by whether people take more money than they deserve based on their performance. Again, however, there is too little theft to perform the intended comparisons. While there is a growing literature in behavioral economics on lying and cheating (e.g. Fischbacher and Föllmi-Heusi, 2013; Gneezy et al., 2018; Kajackaite and Gneezy, 2017), it is unclear to what extent these findings carry over to stealing (Belot and Schröder, 2013; Hermann and Mußhoff, 2019).

for individuals to reconcile pilfering a marker than its monetary equivalent in banknotes with a self-image of being a moral, trustworthy person.[10] Third, our approach mimics workplace theft that often manifests in taking items home from the office, shop, or factory. It also captures the practice of using work resources (the copy machine, envelops, etc.) for one's own purposes.[11]

The paper proceeds as follows. Section 2 presents our theoretical framework and the predictions derived from it. Section 3 introduces the context and design of our experiment. Section 4 presents descriptive statistics as well as our approach to analysis. Our main results and secondary analysis are presented in Section 5. We discuss the interpretation of our results, alternative explanations and the generalizability of our findings in Section 6. Section 7 concludes. Proofs of the theoretical results and relevant tables and figures can be found in the Appendix. The Online Appendix contains details of the experimental procedure (including the instructions), and additional tables and figures.

2 Theoretical framework

This section presents our theoretical model. All proofs can be found in the Appendix. Our approach follows the framework developed by Pierce et al. (2020) to illustrate how loss-framed incentives may affect participants' behavior in multitasking settings. Our rationale for using a framework with multitasking is analogous to that in Pierce et al. (2020): when agents can act on several dimensions, providing loss-framed incentives on one of those dimensions can result in undesired effects across the other dimensions. We preserve the original model's insight about the interaction between loss-framed incentives and multitasking, but we simplify it to a context without uncertainty.[12] We assume that participants' utility under gain-framed incentives depends on the task score s, the amount of theft t, and the effort exerted on the survey z, and can be characterized by

$$v(s+t) + rp(z) - c(s+\alpha z) - \kappa(t), \tag{1}$$

where s and t are expressed in dollars, so that function v measures the utility from combined income. There are two costs functions: $c(s+\alpha z)$ captures the cost of effort to obtain score s and survey completion effort z (α is a scaling parameter that transforms units of effort in survey completion into units of effort in the task), and $\kappa(t)$ represents the moral cost of theft. Finally, the participant derives a moral reward r when she completes the survey adequately,

[10] Consider this example from (Mazar et al., 2008, p.634): "intuition suggests that it is easier to steal a 10-cent pencil from a friend than to steal 10 cents out of the friend's wallet to buy a pencil because the former scenario offers more possibilities to categorize the action in terms that are compatible with friendship (e.g., my friend took a pencil from me once; this is what friends do)."

[11] This behavior is very prevalent and costly for organizations: according to the Association of Certified Fraud Examiners (2016), asset misappropriation is by far the most common form of occupational fraud, and non-cash schemes amount to about a fifth of all cases.

[12] The focus of this paper is to contrast behavior in gain- and loss-framed incentives. This contrast can be examined without explicitly incorporating uncertainty. We therefore abstain from including uncertainty in the formal model as it would add complexity without providing further insight.

and this happens with probability $p(z)$.[13] To mirror the principal-agent problem arising in real organizations, we assume that the experimenter benefits from higher s and z and lower t. Note, however, that the experimenter only directly incentivizes s.[14]

We make standard assumptions with respect to the functions in Equation 1: all of which are twice differentiable in the relevant domain, v and p are increasing and concave (we assume that the participant derives non-pecuniary benefits from helping the experimenter). In addition, we assume that c is increasing and convex, with $c(0) = c'(0) = 0$, and that $\kappa(t)$ has a discontinuity at $t = 0$, so that $\kappa(0) = 0$ but $\kappa(t) = \bar{\kappa}_0 + \bar{\kappa}_1(t)$ for $t > 0$, with $\bar{\kappa}_0 > 0$, $\bar{\kappa}_1(0) = \bar{\kappa}_1'(0) = 0$, and $\bar{\kappa}_1(t)$ increasing and convex. This is justified by the fact that while not stealing has no moral cost, stealing any amount (no matter how small) will have a non-negligible moral cost.[15]

Note that the utility function presented above in Equation 1 represents participants' utility in the *Reward* treatment. Below, and following Pierce et al. (2020), we present an augmented utility function that accounts for the different ways in which loss-framed incentives affect utility. In particular, a participant's utility in the *Penalty* treatment can be captured as

$$u(s, t, z) = v(s + t) + \gamma r p(z) - c(s + \alpha z) - \gamma \kappa(t) - \Lambda[R - s - t]_+,$$

where $[x]_+$ denotes 0 if x is negative and x if x is non-negative. This utility function generalizes the expression from Equation 1 in two dimensions. Firstly, it incorporates *loss aversion*, as for all $\Lambda > 0$ u incorporates a penalty of $\Lambda(R - s - t)$ when a participant's combined income from the task and theft falls below the reference point ($R > s+t$).[16] Secondly, loss-framed incentives can activate the *behavioral spillovers* channel: when $\gamma < 1$, loss-framed incentives alter the utility function by lowering the moral cost of selfish and unethical behavior, making it less costly for participants to refuse to help the experimenter and to steal. Note that $\Lambda = 0$ indicates that the *loss aversion* channel is inactive, and $\gamma = 1$ indicates that the *behavioral spillovers* channel is inactive. If both channels are inactive, that is when when $\Lambda = 0$ and $\gamma = 1$,

[13] Alternatively, $p(z)$ could be interpreted as the utility that the participant derives directly from investing effort z in the completion of the survey due to, for instance, "warm glow" or moral considerations. In this case, r can be interpreted as a scaling parameter.

[14] This is a key feature in our theoretical framework and our experiment. It is designed to mirror the fact that almost every job has elements that benefit/harm the employer, but that the employee is not directly rewarded/punished for.

[15] To assume a discontinuity at 0 is quite natural. For one, there is a discretionary shift in participants' self-perception as being honest when they steal even a small amount (Gilboa et al., 2020) – even though, as stated above, theft of non-monetary items may help preserve one's self-image as honest despite small amounts of theft. Another piece of supporting evidence for this assumption is prospect theory's prediction that individuals overweight small probabilities (Kahneman and Tversky, 1979). That is, even though subjects' probably assumed that they would likely not be caught, they still incur a cost of fearing being caught starting with the theft of a single item. Finally, research in other areas of behavior has shown that people treat zero distinctly differently than positive amount. For instance, not paying subjects for a task can make them be more productive (due to intrinsic motivation) than paying them very small amounts (Gneezy and Rustichini, 2000).

[16] We make the assumption that participants compare their *combined income* of s and t with the reference point. If participants instead only compared their score s with the reference point, then our model would predict that they steal *less* in loss- than in gain-framed incentives. This result arises because participants' utility function would then be $v(s + t) + \gamma r p(z) - c(s + \alpha z) - \gamma \kappa(t) - \Lambda[R - s]_+$ and participants would have an incentive to increase s to avoid the loss aversion penalty. An increase in s, in turn, lowers the marginal utility of $v(s + t)$ because v is concave, which should lower t. Since our results indicate a substantial increase in the amount of theft, we abstain from discussing the case of subjects' only considering their task score.

the utility function simplifies and becomes identical to the one in the *Reward* treatment.

2.1 Comparative statics

This section derives optimal participant behavior in the two treatments. Since behavior under loss-framed incentives is possibly driven by either the *loss aversion* and/or the *behavioral spillover* channels, we will look at each channel's predictions for behavior. We will also derive predictions that allow us to differentiate in the data which of the two channels drives behavior in the *Penalty* treatment. Each participant solves their optimization problem by choosing optimal task score s^*, amount of theft t^*, and survey completion effort z^*. Since there is a discontinuity in the cost function $\kappa(t)$ at $t = 0$, we need to solve the model separately for the cases $t^* = 0$ and $t^* > 0$. Note that the comparative statics derived below to help us to establish whether *loss aversion* or *behavioral spillovers* are driving behavior in the *Penalty* treatment, only hold for the *loss aversion* channel when $s^* + t^* < R$.[17]

Proposition 1. *Comparing the Penalty to the Reward treatment, we obtain the following comparative statics.*

♢ *The task score s^* is higher in the Penalty than in the Reward treatment when the loss aversion channel is activated. When the behavioral spillovers channel is activated, the relative magnitude of s^* is unclear.*

♢ *Theft t^* is higher in the intensive margin (i.e. for $t^* > 0$) in the Penalty than in the Reward treatment, irrespective of which channel becomes activated.*

♢ *The survey completion effort z^* is lower in the Penalty than in the Reward treatment, irrespective of which channel becomes activated.*

The intuition for the result is as follows. For the *loss aversion* channel, as Λ increases, the participant has more incentives to increase both task scores and theft in order to reduce the loss aversion penalty, and this crowds out survey completion effort. For the *behavioral spillovers* channel, as γ decreases, the cost of theft and the reward for completing the survey decrease, and thus theft increases and survey completion effort decreases. Therefore, the direction of change for task scores is ambiguous, as both the marginal utility $v'(s + t)$ and the marginal cost of effort $c'(s + \alpha z)$ are lower.

As there is a discrete jump in the moral cost of stealing (from $\kappa(0) = 0$ to $\bar{\kappa}_0$ for a negligible amount of theft), we need to consider how the extensive margin of theft changes when a participant is in the *Penalty* rather than in the *Reward* treatment.

Proposition 2. *Activating either the loss aversion or the behavioral spillovers channel leads to a (weakly) higher share of participants for whom $t^* > 0$ in the Penalty treatment as compared to the Reward treatment.*

[17] The intuition is simple: loss aversion only matters for behavior as long as people perceive themselves to be in the loss domain with $s^* + t^* < R$. Formally this is so because the term $-\Lambda[R - s - t]_+ = 0$ when $s^* + t^* \geq R$ for any Λ. Therefore, an increase in the intensity of the *loss aversion* channel, represented by an increase in Λ, does not affect a participant's utility function, and thereby their behavior, when $s^* + t^* \geq R$.

The intuition for this result is as follows. Proposition 1 shows that the intensive margin of t^* is higher when moving from gain- to loss-framed incentives for both channels. For this reason, the fixed cost of theft $\bar{\kappa}_0$ becomes less relevant (as compared to the total cost $\kappa(t^*)$). Hence, it becomes optimal for participants to switch from no theft to a positive amount of theft when the influence from either channel is strong enough (when Λ high or γ low enough).

In summary, according to Propositions 1 and 2, we expect (weakly) higher task scores in the *Penalty* than in the *Reward* treatment when the *loss aversion* channel is activated. We do not have a clear predication on task score if the *behavioral spillovers* channel is at work. Both channels imply increased theft (both on the intensive and extensive margins) and a lower level of survey completion in the *Penalty* than in the *Reward* treatment.

2.2 Differentiating between the two channels

As we saw in Propositions 1 and 2, the *loss aversion* and *behavioral spillovers* channels yield similar comparative statics on theft t^* and survey completion effort z^*. They do differ, however, in certain important dimensions. To see this, let us define s^{**}, t^{**}, and z^{**} as the solutions to the problem:

$$(2) \qquad \max_{s,t,z} v(s+t) + \gamma r p(z) - c(s+\alpha z) - \gamma \kappa(t), \qquad \text{s.t. } s+t \geq R.$$

In other words, s^{**}, t^{**}, and z^{**} are the solutions to the maximization problem of an agent who always chooses combined income greater or equal to the reference point R. We expect this to happen when *loss aversion* drives behavior, which is the case when $\Lambda \to +\infty$. The *behavioral spillovers* channel by comparison drives behavior when $\gamma \to 0$.

Proposition 3. *When loss aversion becomes strong (i.e. $\Lambda \to +\infty$), the solutions of the original problem converge to those of Problem (2), i.e. $s^* \to s^{**}, t^* \to t^{**}$, and $z^* \to z^{**}$. As behavioral spillovers becomes strong (i.e. $\gamma \to 0$), scores s^* and survey completion effort z^* converge to 0, and theft $t^* \to +\infty$.*

Proposition 3 shows that the two channels affect participants' behavior differently. When *loss aversion* is driving behavior, participants tend to converge to the solution that they would choose under the restriction $s + t \geq R$. Once they reach that point, further loss aversion (i.e. higher Λ) is irrelevant. By comparison, when *behavioral spillovers* are driving behavior ($\gamma \to 0$), participants have decreasing incentives to complete the survey (as the payoff $rp(z)$ is multiplied by γ), and increasing incentives to steal (as the theft cost $\kappa(t)$ is also multiplied by γ).

2.2.1 Bunching

Loss aversion has been shown to induce *bunching*, which refers to the phenomenon by which a disproportionate amount of individuals place themselves just above or below a certain threshold, e.g. marathon runners attempting to finish below certain "round number" times (Allen et al., 2017), and tax payers reporting income just below a certain threshold in response to differences in marginal tax rates (Kleven, 2016). We formalize bunching following the formal framework developed by Allen et al. (2017) in order to further differentiate between the two channels' predictions. Let $\mathcal{C}$ be a set of cost families $c(\cdot), \kappa(\cdot)$.[18]

Definition 1. *Given set $\mathcal{C}$, we define $\mathcal{C}^+(\delta, x)$ as the set of cost functions in $\mathcal{C}$ such that the agents choose combined income larger than x by at most δ: $\mathcal{C}^+(\delta, x) = \{(c_i, \kappa_i) \in \mathcal{C} : s_i^* + t_i^* \in (x, x+\delta)\}$. Analogously, we define $C^-(\delta, x)$ as those cost functions such that combined income is below x by at most δ: $\mathcal{C}^-(\delta, x) = \{(c_i, \kappa_i) \in \mathcal{C} : s_i^* + t_i^* \in (x-\delta, x)\}$.*

We use the notation $\mathcal{C}^+_{Reward}$ and $\mathcal{C}^+_{Penalty}$ to denote those sets in the respective treatments, and analogously for $\mathcal{C}^-_{Reward}$ and $\mathcal{C}^-_{Penalty}$. With these definitions in place, we can now formalize the concept of bunching in the context of our experiment.

Definition 2. *There is more bunching in the Penalty treatment at x, if and only if there exists a $\delta^* > 0$ such that for any $\delta > 0$ with $\delta \leq \delta^*$: $\mathcal{C}^+_{Reward}(\delta, x) \subset \mathcal{C}^+_{Penalty}(\delta, x)$, and $\mathcal{C}^-_{Reward}(\delta, x) \not\subset \mathcal{C}^-_{Penalty}(\delta, x)$.*

Formally, there is more bunching at R in the *Penalty* treatment than in the *Reward* treatment when the set of cost functions that would generate combined incomes to the right of R is larger in the *Penalty* treatment, and the set of cost functions that would generate combined incomes to the left of R is not larger in the *Penalty* treatment.

Proposition 4. *If the loss aversion channel is activated, we expect more bunching of combined income in the Penalty than in the Reward treatment at $x = R$, and at no other point; whereas if only the behavioral spillovers channel is activated, there is no point at which we expect more bunching in the Penalty than in the Reward treatment.*

In other words, Proposition 4 shows that we expect a shift in the distribution of combined income $s^* + t^*$ from just below R to just above R when the *loss aversion* channel drives behavior, but not when the *behavioral spillovers* channel drives behavior.

3 Context & Design

3.1 Context

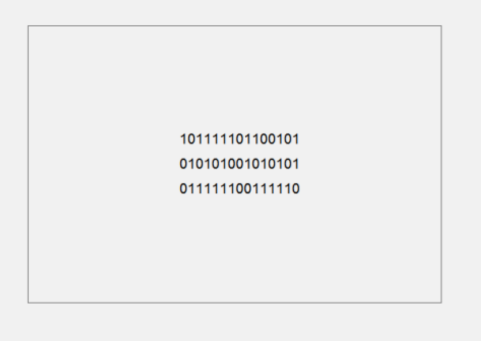
The experiment was conducted at the Erasmus University of Rotterdam in November-December 2014. Participants were recruited using the Online Recruiting System for Economic Experiments (ORSEE). 320 individuals participated in the experiment. To ensure privacy and to minimize any feelings of scrutiny, participants were seated in individual, soundproof cubicles. Each cubicle had a little window in the door, which we covered with paper. The experimenter remained in a different room throughout the experiment. The experiment consisted of a computerized and a pen-and-paper part, the former programmed using the software z-Tree (Fischbacher, 2007). The show-up fee was 2€, and average earnings excluding the show-up fee were 8.6€. Earnings

[18] We follow Allen et al. (2017) in assuming that there can be heterogeneity in cost functions.

were well in line with average student wages at the time. The experiment was conducted in English. The instructions are reproduced in Online Appendix A.2.

3.2 Design

Participants in our experiment worked on a computerized real effort task. Following Abeler et al. (2011), we used a variant of the matrix task that required participants to count the number of zeros in matrices of randomly ordered zeros and ones (see Figure 1 for an example). The matrices consisted of 3 rows and 15 columns, and participants had 10 seconds per round to count how many zeros they contained.



```
101111101100101
010101001010101
011111100111110
```

Figure 1: Screenshot of the computerized matrix task

After 5 unpaid practice rounds, participants completed 100 payment-relevant rounds, seeing a new matrix in each round. Participants received immediate feedback on whether their answer was correct after each round. At no point during the task did they learn their aggregate score or their relative performance compared to other participants. We chose a task that was tedious and boring and without any higher purpose, in order to minimize participants' intrinsic motivation to do well.

We used a between-subject design, randomly assigning participants to either the *Reward* or the *Penalty* treatment. The two treatments only differed from each other in the incentive schemes we used to translate participants' performance in the matrix task into earnings. In the *Reward* treatment, participants were told they would receive 10 cents (0.1€) for every matrix they solved correctly, and would be paid the amount they earned at the end of the experiment. A perfect score (100 correct answers) thus earned participants a payoff of 10€. In the *Penalty* treatment, participants received 10€ upfront and were told they would lose 10 cents for every incorrect answer, and would have to return the total amount lost at the end. Following the procedure in Levitt et al. (2016), participants in the *Penalty* treatment signed the following receipt upon receiving the 10€ banknote at the beginning of the experiment: *"I hereby confirm the receipt of 10€ before the start of the experiment. These are mine and belong to me."* Note that the two payment schemes were payoff-equivalent: the same task performance lead to the same actual earnings, but they were framed either as gains or losses depending on the treatment. Treatments were made salient by frequent reminders. After each wrong answer participants in the *Penalty* treatment saw a red panel on the screen with the message “YOU LOST MONEY!”. In order to move on to the next round, participants then had to click a button saying “I LOST MONEY!”. On the flip side, participants in the *Reward* treatment saw green “YOU EARNED MONEY!” panels after each correct answer, and had to click a button saying “I EARNED MONEY!” in order to proceed to the next round.

After completing the required 100 rounds of the matrix task, participants had to fill out an obligatory questionnaire. The questionnaire was placed on participants' desks in paper format, and contained non-incentivized questions on demographics (name, student number, age, gender, year of study, major), guessed task performance, whether they would recommend participation in the experiment to their friends, and whether they would want to take part in another round of the same experiment within the following weeks.[19] Participants were also asked to rate on a seven-point scale how hard they had worked, how happy they felt, how much fun they found the task to be, and how fair/adequate they thought the payment scheme was.

Finally, after participants had completed the obligatory questionnaire, they were invited to fill out another survey. We informed participants that participation in this additional survey was voluntary and that there would be no reward or punishment associated to completing it. The survey was part of an unrelated research project and focused on the topic of flexible work arrangements.[20] This additional survey was also in paper format and included multiple choice questions, open-ended questions, and free text fields that elicited suggestions on how to improve the survey. Participants were asked to complete all questions and text fields as only complete surveys could be evaluated by the experimenter. We use survey completion as our measure of uncompensated helping because it captures a participant's willingness to exert voluntary effort that benefits the employer (in our study, the researcher conducting the experiment) with little or no benefit to the participants themselves (Bradler and Neckermann, 2016).[21]

We use a new experimental paradigm to measure theft in the laboratory. We placed a large box of office supplies on the desk in each participants' cubicle, and recorded whether any items were missing from the container after the experiment. Each box contained 3 pencil sharpeners (2.50€) and 10 each of the following items: pencils (0.1€), pens (0.2€), erasers (0.3€), post-it notes (0.5€), correction rollers (0.75€), fine liners red (0.8€), fine liners blue (0.8€), yellow markers (1€).[22] The items were all mixed together, making it impossible to determine simply by glancing at the box and without actually counting the items to see whether anything was missing: participants could therefore reasonably assume that the theft of a small number of items would go unnoticed. The experimental instructions explicitly brought the participants'

[19] Because only a randomly selected small subset of those who signed up for another round were actually invited back to participate in the extra session, we can not use the actual show-up decision as a measure of retention.

[20] Participants found the voluntary survey under the obligatory questionnaire on their desk. Both surveys were handed to the experimenter at the end of the experiment when subjects left their cubicles to receive their payment. See Online Appendix A.4 for a copy of the survey.

[21] Relatedly, Danilov and Vogelsang (2016) show that prosocial behavior can manifest itself in the lab as time invested in order to benefit another participant.

[22] The numbers in brackets show the approximate monetary value of each item. Source: *hema.nl*, website of a large Dutch retailer for office supplies. Figure A2 in Online Appendix A.1 shows the boxes and the elements they contained.

attention to the box when describing the obligatory questionnaire: *"You find it at the top of your desk under the container with the pencils. [...] There are also pencils and other material provided on your desk."* There was no mention, however, of taking office supplies home: it was neither encouraged nor forbidden. In our opinion, this method provides a natural and inconspicuous way to measure stealing.[23]

We end this section by pointing out some important features of our design. First, the participants were presented with the box of office supplies and were required to read the full set of instructions before beginning the real effort task. As such, we find it unlikely that the participants made their choices regarding task effort, survey effort and theft in a strict sequence; rather, we believe that a simultaneous choice model provides a more accurate approximation of participant behavior in the experiment.

Second, all the decisions of interest (signing up for another round of the same experiment; filling out the voluntary survey; stealing office supplies) happened *before* the participants learned their aggregate score and received/returned money. That is, the participants did not find out about any possible discrepancy between their expected and actual scores until after they had made all of their choices.

Third, participants in our experiment *incurred* but did not actually *realize* their losses from task earnings before making the decision to steal: they did not physically give up part of their endowment until after they had left their cubicles at the end of the experiment. We therefore assume that these "paper losses" (Imas, 2016) from the task were mentally bracketed together with the pecuniary gains from theft (see also Footnote 16). This implies that we assume that the participants considered their combined income from task and stealing when assessing whether they had incurred losses or gains.

We take the reference point in the *Penalty* treatment as the initial endowment of 10€ (that corresponds therefore to the initial status quo). This follows a well-established literature that considers that the reference point is exogenously determined by the endowment or status quo (Masatlioglu and Ok, 2014; Ortoleva, 2010; Riella and Teper, 2014; Tversky and Kahneman, 1991).[24]

[23] It is worth noting that even though the experimenter did not enter the cubicle until after the participant had left and participants were aware of this, the cubicle number was used to determine payment at the end. Hence, while participants could not be "caught red handed", participants may have perceived it possible that theft would be discovered and linked to their name. We therefore expect the prevalence of theft in our experiment to be a conservative estimate compared to perfectly anonymous situations where the risk of being exposed is eliminated.

[24] A more recent literature allows the reference point to be endogenous, but still assumes that the reference point to a large extent depends on the initial endowment (Barbos, 2010; De Giorgi and Post, 2011; Guney et al., 2018; Koszegi, 2006; Maltz, 2020; Ok et al., 2015). Note that our setting can accommodate the existence of loss aversion in the *Reward* treatment as long as participant reference point is close to their initial endowment of zero or at least below 5€. In both of these cases there would be no loss aversion penalty.

4 Data & Analysis

4.1 Descriptive statistics

Our sample consists of 320 participants of whom 161 were exposed to loss-framed incentives. Our treatment groups are balanced in terms of demographic characteristics (see Table 3 in the Appendix). Table 1 provides the summary statistics of the demographic variables, the performance in the task, and participants' elicited opinions about the experiment.

Table 1: Summary statistics

Variable (scale)	Mean	Std. Dev.	Min.	Max.	N
Demographic variables					
Male (0/1)	0.6	0.49	0	1	320
Age	21.9	3.3	17	51	307
Year of study	3.24	1.57	1	6	319
Econ student (0/1)	0.69	0.46	0	1	320
Performance and effort					
Task performance (0-100)	85.97	10.56	43	100	320
Guessed performance (0-100)	80.21	15.88	20	100	320
Self-reported effort (1-7)	5.74	1.42	1	7	307
Evaluating the experiment					
Happy (1-7)	4.98	1.21	1	7	319
Fun (1-7)	4.12	1.62	1	7	320
Fair (1-7)	5.46	1.37	1	7	320
Return	0.91	0.28	0	1	320
Refer friends (0/1)	0.85	0.36	0	1	319

60% of our participants are male and 69% are Economics students. The average age is 21.9 years. Overall, performance in the matrix task is rather high (the mean score is 85.97 out of 100, varying between 43 and 100), and participants are quite accurate in guessing the number of questions they solved correctly: the raw correlation between actual and guessed task performance is 0.85.[25] Participants' self-reported effort provision is also on the high side, with a mean of 5.74 on a scale from 1 to 7. Even though participants do not find the experiment particularly fun (on average, they rate it 4.12 on a scale from 1 to 7), they consider the payment adequate/fair (a mean rating of 5.46), and 91% (85%) would be willing to return for another round of the same experiment (recommend participating in the experiment to their friends). Note that participants may have expressed these relatively positive attitudes about the experiment because the questionnaire was not anonymous.[26]

4.2 Approach to analysis

Our main analysis compares participant behavior between the two treatment conditions. Our primary outcome variables are constructed from the real effort-based and payment-relevant

[25] Remember that participants received immediate feedback after submitting each answer, but were not told their total score until the payment stage at the very end of thee experiment.

[26] Answers from the questionnaire are missing for some students: we observe age, reported happiness and willingness to refer friends for 319 and age and effort for 307 out of 320 participants. The experimenter ensured that all questionnaires contained student names and ID numbers.

measures of task effort, survey effort, and theft that we collected in the experiment. In our main analysis, we use task score (the number of correct answers in the matrix task) as our proxy for effort spent on the task.[27] We analyze two measures of theft: the binary decision to steal and the estimated value of the items stolen. We present these two measures to maintain comparability with the existing literature on theft in experimental economics (e.g., Gravert, 2013; Hermann and Mußhoff, 2019) and to test our theoretical predictions with regards to the intensive and extensive margins of theft. Our proxy for survey effort is the binary measure of completing the voluntary survey.[28]

Our secondary analysis explores the mechanisms that lead to differences in behavior between the treatments. We do so by studying the following secondary outcome measures: participants' combined income (defined as the sum of participants' task-related payments plus the estimated monetary value of the items they have stolen – if applicable) and their satisfaction with the experiment (as measured by the first principal component of the five variables from the obligatory questionnaire on the participants' experience: how happy they felt, how fun the task was, how fair the compensation was, whether they were willing to return for another round and whether they were willing to recommend the experiment to their friends). We treat the results based on the latter measure as suggestive, as they could be subject to experimenter demand effects or social desirability bias.

In the discussion, we consider alternative measures of task effort (guessed task performance and self-reported effort from the questionnaire; time spent per question on the task). We also explore how robust our results are to alternative ways of defining and measuring theft.

Throughout the section, we use two-sided t-tests with unequal variances to compare the means of continuous outcome variables. For comparing binary outcome variables we use two-sample chi-squared tests of proportions. In addition to conventional p-values, we also report randomization inference-based p-values for all comparisons, and the p-value from a Westfall-Young joint test of statistical significance (Young, 2019) for our four main outcome variables (task performance, two measures of theft and survey completion).

5 Results

This section shows the empirical results from our experiment. Sections 5.1-5.3 present our main analyses that test the predictions of our theoretical model regarding differences in task effort, survey effort and theft between the *Reward* and the *Penalty* treatments. Sections 5.4 and 5.5 present exploratory analyses of the channels that may explain the differences in participants' response to gain- vs. loss-framed incentives by studying participants' combined income and satisfaction in the experiment.

[27] We also measured time spent on each matrix, the result of which will be discussed in Section 6.

[28] The instructions for the voluntary survey informed participants that "Only completed surveys can be evaluated" suggesting completion is the most welfare-relevant measure. We thank an anonymous reviewer for making this recommendation.

5.1 Main analysis: task score

We begin our analysis by comparing how participants performed in the real-effort task under the two treatment conditions. Recall that Proposition 1 predicts (weakly) higher task scores in the *Penalty* than in the *Reward* treatment when the *loss aversion* channel is active, whereas the direction of change resulting from the *behavioral spillovers* channel is ambiguous.

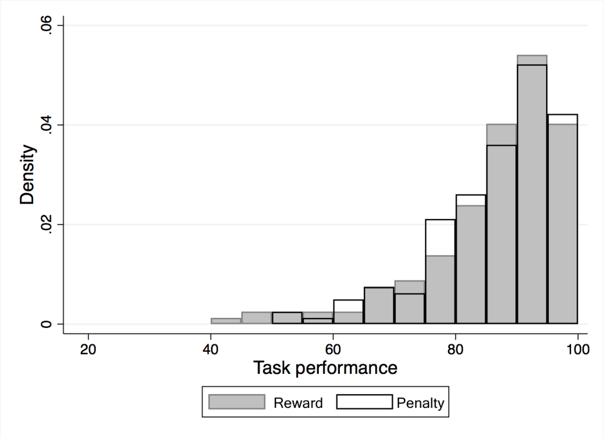


Figure 2: Distribution of task scores by treatment

Figure 2 compares the distribution of total scores in the matrix task. In our study, there is no significant difference in performance between those who experience gain- vs. loss-framed incentives: the group means are 85.7 and 86.2 in the *Reward* and *Penalty* treatment, respectively; a difference that corresponds to less than 1% of the mean in the *Reward* treatment, or approx. 5% of the pooled standard deviation. A t-test of the difference in means yields a p-value of 0.661. Following the approach in Young (2019), we obtain a randomization inference-based p-value of 0.652.

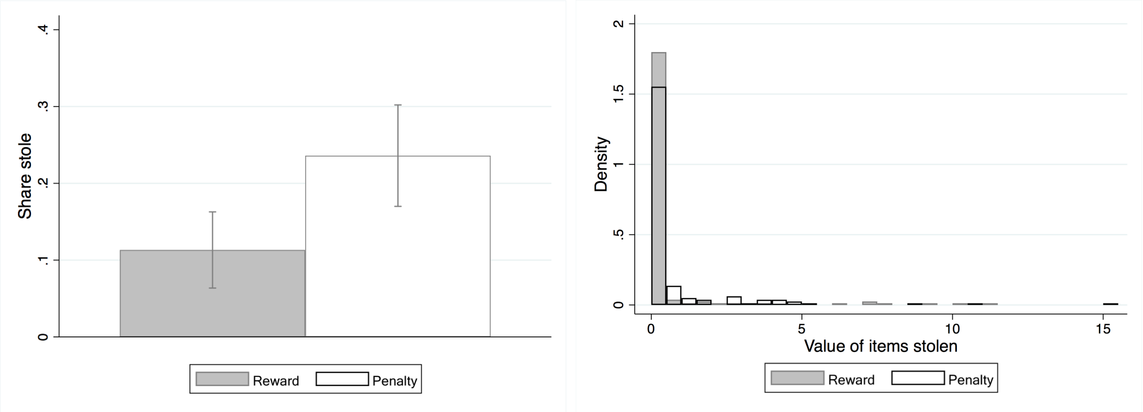
5.2 Main analysis: theft

We continue by comparing the intensive and extensive margin of theft in gain- and loss-framed incentives. Propositions 1 and 2 predict an increase along both margins when either the *loss aversion* or the *behavioral spillovers* channel is activated.

Our data show a clear impact of loss-framed incentives on theft. As can be seen in Figure 3a, participants assigned to the *Penalty* treatment are substantially more likely to steal: the share of participants who take at least one item from the box of office supplies is more than twice as high in the *Penalty* than in the *Reward* treatment (11.3% in the *Reward* treatment and 23.6% in the *Penalty* treatment; a two-sample test of proportions yields a p-value of 0.004, and the randomization inference-based p-value is 0.003). That is, while in the *Reward* treatment only 18 out of the 159 participants steal anything, the corresponding number is 38 out of 161 in the *Penalty* treatment.

The treatment has an effect on the intensive margin of theft as well. The mean value of items stolen (including zeros) is 44% higher in the *Penalty* than in the *Reward* treatment: it is

0.47€ in the *Reward* and 0.67€ in the *Penalty* treatment. The p-value from a t-test comparing the mean value stolen between treatments is 0.336 (the randomization inference-based p-value is 0.338). Figure 3b presents the distribution of the value of stolen items and suggests that the *Penalty* treatment disproportionately induces "small" theft.



(a) Share of participants who stole

(b) Distribution of the value of items stolen

Figure 3: Theft by treatment

5.3 Main analysis: survey completion

This subsection looks at the effect of loss-framed incentives on survey completion, a voluntary act of service by the participant towards the experimenter. Proposition 1 predicts lower survey completion rates in loss- than in gain-framed incentives when either channel is activated.

Figure 4 compares the share of participants who completed the voluntary survey between the two treatments. In the *Reward* treatment 18.1% of participants complete the survey, whereas only 14.5% do so in the *Penalty* treatment. This represents a reduction of 3.6 percentage points or 21.5%, which is not statistically significantly different from zero (p-value from a two-sample test of proportions is 0.387; randomization inference-based p-value is 0.364).

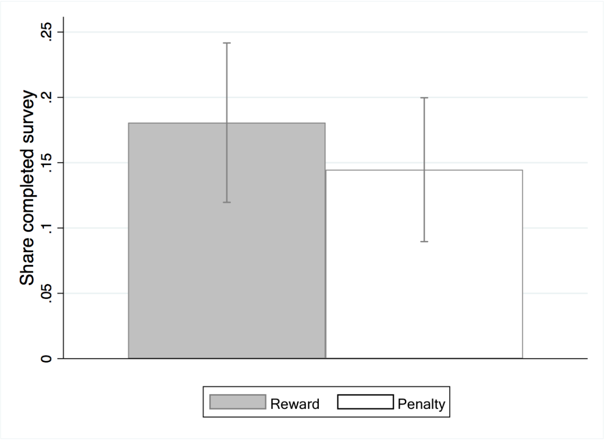


Figure 4: The impact of loss-framed incentives on survey completion

Using the randomization inference-based approach outlined in Young (2019), we can conduct a joint test of the sharp null hypothesis that the treatment had no effect on any of our main outcomes (task score, share who stole and value stolen, survey completion). This test yields a p-value of 0.017, so we can reject the hypothesis that our treatments were completely irrelevant for participant behavior.

5.4 Secondary analysis: combined income

Our main analysis has established that participants respond to loss-framed incentives. As we discussed in Section 2, we however need to go beyond comparative statics to explore which channel drives this response. Recall that Proposition 4 predicts more bunching of combined income (from task earnings and theft) at the reference point in the *Penalty* than in the *Reward* treatment when the *loss aversion* channel is activated. This prediction is unique to the *loss aversion* channel: if only the *behavioral spillovers* channel is active, we do not expect to see more bunching in the *Penalty* treatment at any point; instead, this channel predicts a universal shift to higher theft.

Figure 5 shows the difference in kernel density estimates of combined income between the *Penalty* and *Reward* treatments (we estimated the kernel densities using .5 half-width and 60 points). As can be readily observed from the graph, there is a sharp decline in the difference of densities before 10€, and a sharp increase after 10€.[29] We interpret this as visual evidence for bunching in combined income around 10€, i.e. a shift of mass in the *Penalty* treatment (as compared to the *Reward* treatment) from the left to 10€ to the right of 10€.

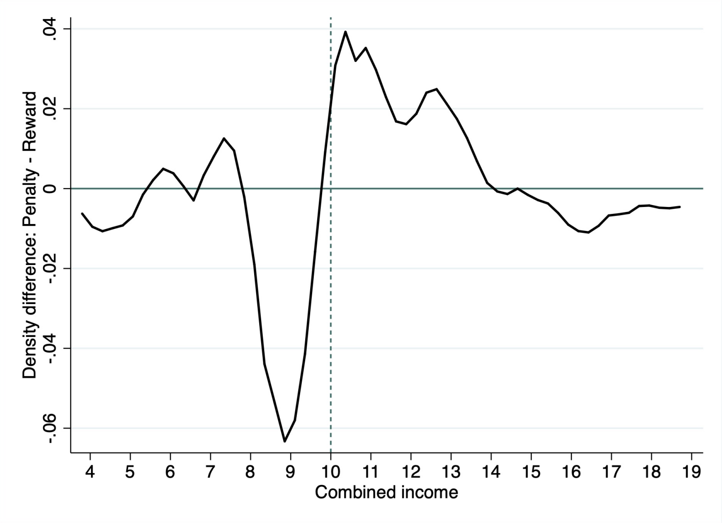


Figure 5: Difference in kernel density estimates between *Penalty* and *Reward* treatments.

We also test for bunching in combined income more formally. Table 2 presents estimated marginal effects from probit models testing whether participants' combined income is more likely

[29] To be precise, the difference of densities crosses the 0-horizontal axis at approximately $s + t \approx 9.8$ rather than 10. We attribute this to error (both purely sampling error, as well as errors in the participant's estimations of their own scores and the value of their theft).

to be above a certain threshold in the *Penalty* than in the *Reward* treatment.[30] In columns (1)-(3) this threshold is the reference point, 10€, whereas columns (4) and (5) present results from placebo tests repeating the exercise at alternative thresholds of 9€ and 11€ respectively. Columns (1), (4) and (5) present results from estimations that include the full sample of participants, whereas columns (2) and (3) restrict the sample to observations that fall within a smaller window around the reference point (specifically, we consider windows of ±1 and ±0.75, respectively).

Table 2: The effect of loss framed incentives on the likelihood of combined income exceeding threshold

	(1)	(2)	(3)	(4)	(5)
Threshold	10	10	10	9	11
Window	None	±1	±0.75	None	None
Penalty treatment	0.093**	0.131**	0.150**	0.043	0.031
	(0.036)	(0.055)	(0.071)	(0.055)	(0.029)
	[.006]	[.011]	[.020]	[.458]	[.290]
N	320	143	110	320	320

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

The table presents estimated marginal effects at the mean from probit models, where the dependent variable is an indicator for participant's combined income (from task earning and theft) exceeding a certain threshold, and the independent variable is an indicator for being in the *Penalty* treatment. The column headers display the specific threshold used in the model presented in each column (10 for columns 1-3, 9 for column 4, 11 for column 5). Column headers also specify whether we have restricted our analysis to only include observations from a specific window around the threshold (columns 1, 4 and 5 present results from models that apply no such restrictions, whereas columns 2 and 3 include observations from windows of ±1 and ±0.75 around the threshold, respectively.). Standard errors in parentheses; * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Randomization inference-based p-values in brackets.

Column (1) shows that the *Penalty* treatment increases the likelihood that participants' combined income falls above 10€ by 9.3 percentage points on average, a difference that is statistically significantly different from zero at the 0.01 level. Columns (2) and (3) confirm that this shift in combined income from below to above 10€ happens close to the reference point: once we restrict our attention to relatively narrow windows around the reference point, we still detect a statistically significantly higher likelihood in the *Penalty* treatment that the combined income exceeds 10€, and the estimated effect size is similar across the three specifications (the effect increases as the window becomes smaller). Importantly, results from the last two columns suggest that the shift happens at around 10€ and not at other values: participants in the *Penalty* treatment are no more likely to earn combined incomes higher than 9€ (column (4))

[30] The empirical literature on bunching (Allen et al., 2017; Chetty et al., 2011; Kleven, 2016) usually considers contexts in which the counterfactual is not observed and must be estimated (for example, using local polynomials around the reference point). In a second step, these papers then compare the counterfactual with the actual data. Because we run an experiment, our data already contains a counterfactual distribution, allowing us to test for bunching with a simple probit regression around the reference point.

or 11€ (column (5)) than participants in the *Reward* treatment.

5.5 Secondary analysis: participant experience

We also make use of participants' answers in the obligatory questionnaire to explore which of the two channels drive behavior in response to loss-framed incentives. In particular, we ask whether we can detect any sign of participants in the *Penalty* treatment experiencing negative emotions, feelings of unfair treatment or lower satisfaction compared to those in the *Reward* treatment – emotions and perceptions that could activate the *behavioral spillovers* channel.

To this end, we analyze the answers to all five questions in the obligatory questionnaire that pertain to participants' experience in the experiment: how happy they feel, how fun the task was, how fair the compensation was, whether they would be willing to return for another round of the same experiment and whether they would be willing to recommend the experiment to their friends. As the answers to the five questions are highly correlated, we summarize them in a "participant satisfaction" index that corresponds to the first principal component of the five variables, rather than performing five individual comparisons.[31] This index ranges from -6.2 to 2.4, with a mean of zero (by construction) and standard deviation of 1.5. According to this index, there is no meaningful difference between participants' experience between the two treatments: mean participants' satisfaction is slightly lower in the *Penalty* than in the *Reward* treatment, but the difference is small in size (0.107 SD) and not statistically significantly different from zero (p-value 0.339, randomization inference-based p-value 0.343). Analyzing the answers to each of the five survey questions separately confirms this conclusion.

6 Discussion

In this section, we offer our interpretation of the empirical results, consider alternative explanations, and discuss the generalizability of our findings.

6.1 Interpretation of results

We start by considering our first main outcome measure: participants' score on the real-effort matrix task. Our results show a much smaller (and not statistically significant) effect of loss-framed incentives on performance than a number of related laboratory experiments. Comparing a gain with a loss treatment, Imas et al. (2017) report a difference of 0.4 SD in the mean number of tasks completed, and Armantier and Boly (2015) find a difference of around 0.3 SD in mean earnings.[32] We argue that the lack of a clear effect of loss-framed incentives on performance in our experiment is unlikely to be driven by low statistical power: our study was sufficiently

[31] Online Appendix B contains tables and figures on the five different variables.

[32] Note, however, that a number of other studies have only found small, insignificant, or marginally significant positive effects of loss-framed incentives on task performance (Brooks et al., 2012; De Quidt et al., 2017; DellaVigna and Pope, 2017; Grolleau et al., 2016; Hong et al., 2015). For a detailed review of the literature, please refer to De Quidt et al. (2017).

powered to detect an effect of similar size as observed in the studies mentioned above.[33] Nor do we think it is attributable to an insensitivity of the task to incentives: Gall et al. (2016) presents evidence that performance in the matrix task is responsive to incentives.

The results are rather consistent with participants reacting to loss-framed incentives through an increase in theft rather than an increase in task performance, a finding that is compatible with both the *loss aversion* and the *behavioral spillovers* channel. Recall that Proposition 1 predicts an ambiguous effect of loss-framed incentives on scores for the *behavioral spillovers* channel. Moreover, the *loss aversion* channel implies *weakly* higher task scores in the *Penalty* than in the *Reward* treatment. In particular, we expect higher task scores only for those participants whose optimal combined income would remain below their reference point and are thus in the loss domain. Since the *loss aversion* channel also predicts an increase in both the intensive and extensive margin of theft (Propositions 1 and 2), we expect more participants in the *Penalty* treatment to end up with combined income above their reference point due to theft. This in turn implies that there is no additional incentive for these participants to work harder on the task than their peers in the *Reward* treatment.

We next turn to our findings regarding theft. Our results are broadly consistent with our model that predicts an increase in theft according to both channels, along the intensive (Proposition 1) as well as the extensive margin (Proposition 2). We find a substantial (109%) and statistically significant increase in the share of participants who steal, and a moderate-sized (44%) and statistically insignificant increase in the average value of items stolen.

Our findings regarding our final main outcome - survey completion - suggest that in our setting loss-framed incentives have a weaker impact on voluntary helping behavior than on theft. The share of participants filling out the voluntary survey is lower in the *Penalty* than in the *Reward* treatment (21.5%), but the difference is not statistically significantly different from zero. This slight decrease is broadly consistent with our model (Proposition 1) that predicts a decrease in survey completion in loss-framed incentives for both channels.

Our secondary analysis provides suggestive evidence that the differences we observe between the two treatment conditions are primarily attributable to the activation of the *loss aversion* channel. First, we presented evidence for bunching in combined income around participants' reference point in the *Penalty* treatment. According to Proposition 4, this behavior is consistent with the *loss aversion* but not with the *behavioral spillovers* channel.[34] Second, answers in the obligatory questionnaire show that loss-framed incentives did not cause participants to express animosity towards the experimenter or question the fairness of their payment, suggesting that participants did not experience the negative emotions required to activate the *behavioral spillovers* channel. Given the strength of our framing intervention with the frequent reminders

[33] We calculated our minimum detectable effect size to be approximately a third of a standard deviation, assuming a significance level of 5% and power of 80%, using a t-test to compare group means.

[34] We would like to note that the *behavioral spillovers* channel predicts a shift of theft to the right (and an ambiguous change in task scores). The reader might wonder whether the shift in mass from below 10 to above 10 in the distribution of combined income could be accounted for by certain parameters within the *behavioral spillovers* channel. However, only a very narrow set of parameters would be able to generate such a shift: γ would have to be small enough to generate the shift in behavior around 10, but not too small, or the entire distribution would move entirely beyond 10 (Proposition 3).

about money lost, one might perhaps be surprised by the absence of any effect on self-reported satisfaction measures. At the same time, Brownback and Sadoff (2019) and De Quidt (2017) also find no effect of loss-framed incentives on subjective well-being and stress levels.

6.2 Alternative explanations

This section discusses alternative explanations that may account for our findings. First, we consider whether the reason we did not observe a more pronounced response to loss-framed incentives in terms of task performance is because actual task scores are not accurate measures of participant effort. We do so by examining alternative measures collected in our experiment: guessed task scores and self-reported levels of effort provision from the obligatory questionnaire, and the number of seconds spent on solving each matrix in the task. Mean *guessed* task score is slightly higher in the *Penalty* than in the *Reward* treatment (81.5 vs. 78.9, p-value = 0.131, randomization inference-based p-value 0.131). Self-reported effort provision is also somewhat higher in the *Penalty* than in the *Reward* treatment, with respective group means of 5.9 and 5.6 (p-value = 0.026, randomization inference-based p-value 0.028). However, this higher self-reported effort is not reflected in a detectable difference in the amount of time participants spent on the task per round before submitting their answers.[35] In sum, using alternative proxies for measuring task effort does not change our conclusion regarding a lack of clear performance impact of loss-framed incentives.

Second, one might wonder whether our measure of theft overstates participants' true intentions to steal. In our analysis, we treat items missing from the box of office supplies as a sign of theft. This interpretation assumes that all missing items were taken on purpose by participants. It could, however, be true that participants simply forget to return a pen or pencil to the box after they used them to fill out the questionnaire. We find little evidence for such unintended theft. Table 4 in the Appendix provides more detailed information on theft, displaying the number of items stolen in each category (pens, pencils, markers, etc.). We find that pens and pencils were among the less popular items pocketed. In additional unreported analyses, we find that conditional on stealing, the vast majority of people take something else or more than just a single pen or pencil: there are only 4 instances where a participant takes nothing but a pen/pencil, and these four cases are equally divided between the two treatments.

6.3 Generalizability

Given that our results were obtained in a laboratory environment with a student sample, it is important to discuss to what extent our finding generalize to employee behavior in organizations.

In particular, one may wonder whether characteristics of the experiment such as its artificial environment and overt nature might affect participant behavior, especially the decisions to steal and to help. We aimed to minimize the level of scrutiny participants experienced by seating

[35] Figures 6 and 7 in the Appendix show the distribution of guessed tasks scores and self-reported effort provision by treatment, while Figure 8 (also in the Appendix) compares the time spent on the calculation task per round by treatment.

them in individual sound-proof and closed cubicles, and we operationalized theft and helping in subtle ways that closely approximate the temptation of asset misappropriation and the moral obligation for organizational citizenship behaviors that employees might experience at work.

Other aspects of the experimental environment, such as the nature of the task, the way the treatment was implemented, the stakes, or the consequences of the theft may not approximate conditions in real organizations perfectly. While our real-effort task is certainly artificial, matrix tasks like ours have been used in many studies to mimic tedious jobs that require concentration (e.g. Abeler et al., 2011). Furthermore, while the stakes in our experiment were not as high as monthly salaries, they were meaningful to our student participants who exerted considerable effort on the task in order to make money. Admittedly, tasks and stakes such as the ones typically used in laboratory study are far from perfect representations of situations outside of the laboratory. As such, we caution against extrapolating the *level* of theft we observed to other environments.[36]

Another potential threat to external validity relates to study populations. We obtained our results among students at a Dutch university. Considering that the typical student is on their way to becoming an employee, one might hope that our findings extrapolate to college-educated Western employees. A large body of research has shown that loss aversion is an important driver of behavior across many populations and domains (Barberis, 2013; Ruggeri et al., 2020). We also find it encouraging that our results are largely in line with those of other studies, using other subject pools, on loss aversion and unethical behavior (Cameron and Miller, 2009; Grolleau et al., 2016; Kern and Chugh, 2009; Pettit et al., 2016; Schindler and Pfattheicher, 2017; Shalvi et al., 2011); and loss aversion and multitasking (Pierce et al., 2020; Rubin et al., 2018).

Still, our experiment was one-off and of short duration. Students neither had prior experience with the task nor were they in an ongoing relationship with their employer, the experimenter. One could easily imagine that any existing hostility between employees and management might be amplified or interact with the institution of loss-framed incentives. Further, our study is not able to address how theft as well as helping and retention would be affected over a longer time period. One might hypothesize that the effects could wear off. Encouragingly, Levitt et al. (2016) and Brownback and Sadoff (2019) study loss-framed incentives over the course of an academic year and do not find any deterioration in the effects that they document.

In sum, we are relatively optimistic about the generalizability of the finding that loss-framed incentives might induce theft or other, possibly undesirable, side-effects as employees attempt to minimize possible losses. We are less certain, however, that we would not find evidence in support of the behavioral spillover channel in real organizations. A number of factors, such as an ongoing employer-employee relationship and communication between employees, might

[36] It is unclear whether our treatment intervention with its frequent reminders was more or less strong compared to how loss-framed incentives are implemented in the field. We would conjecture that there are fewer reminders in a typical field setting, but that loss-framed incentives in the field are not less salient due to the stakes as well as the rarity of working under such incentives. In addition, there was no opportunity for interaction or communication between participants in our experiment. Communication and interactions are, however, important factors in real workplaces. One could imagine that employees might be even more inclined to steal if they see others doing so. Similarly, theft might be even higher if employees feel annoyed by the structure of the incentives and can share this sentiment with others.

make it more likely that loss-framed incentives induce negative behavioral spillovers outside of a controlled laboratory setting.

7 Conclusion

Our experiment extends the study of loss-framed incentives beyond their impact on employees' effort and performance (actual or self-reported) to include outcomes such as stealing and helping. We find that loss-framed incentives double the proportion of participants who steal and increase the value of items stolen by 44 percent compared to gain-framed incentives. There is also a small, not statistically significant reduction in participants' willingness to complete a voluntary survey, our proxy for uncompensated helping. Our results are consistent with the explanation that loss aversion is driving these behaviors.

Our study has important implications for management. In our experiment, loss-framed incentives backfired: they did not meaningfully increase performance, but they did increase theft. As such, we caution against the use of loss-framed incentives in organizations where multiple strategies are available for employees to reduce their losses. Furthermore, the fact that we observed a relatively small and insignificant reduction in voluntary helping behavior in response to loss-framed incentives might be an artefact of the experimental nature of our study. Managers in real firms might be more likely to see negative behavioral spillovers from loss-framed incentives and therefore a drop in voluntary helping.

These pieces of evidence may help us understand why loss-framed incentives are used so rarely in organizations despite the fact that an increasing number of experimental studies advertise their effectiveness. Future research will need to delve deeper into the study of various incentives schemes in complex work environments to improve our understanding of the conditions and contextual factors that inhibit or promote the overall effectiveness of rewards beyond a narrowly defined measure of output.

Appendix

Proofs of the theoretical results

Proof of Proposition 1. We begin begin solving the optimization problem for the agent. We need to consider the First Order Conditions for the cases $t^* > 0$ and $t^* = 0$. We start with the case in which theft is strictly positive ($t^* > 0$), and take the first order conditions with respect to s, t and z:

$$v'(s+t) - c'(s+\alpha z) + \Lambda \mathbb{1}_{R>s+t} = 0, \tag{3}$$

$$v'(s+t) - \gamma\kappa'(t) + \Lambda \mathbb{1}_{R>s+t} = 0, \tag{4}$$

$$\gamma r p'(z) - \alpha c'(s+\alpha z) = 0. \tag{5}$$

From Equation 5, we obtain

$$c'(s+\alpha z) - \frac{\gamma r}{\alpha} p'(z) = 0. \tag{6}$$

From Equations 3 and 4 we derive $c'(s+\alpha z) = \gamma\kappa'(t)$, and using that in Equation 4, we obtain

$$\kappa'(t) - \frac{r}{\alpha} p'(z) = 0. \tag{7}$$

Equations 4, 6 and 7 jointly characterize the optimal s^*, t^* and z^* for the case $t^* > 0$. For the problem with $t^* = 0$, the first order conditions are:

$$\begin{align} v'(s) - c'(s+\alpha z) + \Lambda \mathbb{1}_{R>s} =& 0, \tag{8} \\ \gamma r p'(z) - \alpha c'(s+\alpha z) =& 0. \tag{9} \end{align}$$

In what follows, for notational simplicity, we will write functions without their arguments in the following calculations: for example, v'' instead of $v''(s+t)$. Recall that $v', p', c', \kappa' > 0$, $v'', p'' < 0$ and $c'', \kappa'' > 0$. For the case $t^* = 0$, the solutions are characterized by Equations 8 and 9. We can use the Implicit Function Theorem to obtain the comparative statics for s^* and z^* with respect to Λ and γ. If we define function G using Equations 8 and 9, then we compute the Jacobian matrix with respect to s, z, and with respect to Λ, γ, respectively:

$$J_{s,z} = \begin{bmatrix} v'' - c'' & -\alpha c'' \\ -\alpha c'' & \gamma r p'' - \alpha^2 c'' \end{bmatrix}, \qquad J_{\Lambda,\gamma} = \begin{bmatrix} \mathbb{1}_{R>s} & 0, \\ 0 & rp' \end{bmatrix}.$$

Therefore, from the Implicit Function Theorem, we have that the matrix of comparative statics for s^* and z^* with respect to Λ and γ is given by:

$$\begin{bmatrix} \frac{\partial s^*}{\partial \Lambda} & \frac{\partial s^*}{\partial \gamma} \\ \frac{\partial z^*}{\partial \Lambda} & \frac{\partial z^*}{\partial \gamma} \end{bmatrix} = -J_{s,z}^{-1} \times J_{\Lambda,\gamma} = -\frac{1}{det(J_{s,z})} \begin{bmatrix} \gamma r p'' - \alpha^2 c'' & \alpha c'' \\ \alpha c'' & v'' - c'' \end{bmatrix} \times \begin{bmatrix} \mathbb{1}_{R>s} & 0 \\ 0 & rp' \end{bmatrix} =$$

$$= -\frac{1}{det(J_{s,z})} \begin{bmatrix} (\gamma r p'' - \alpha^2 c'') \cdot \mathbb{1}_{R>s} & \alpha r p' c'' \\ \alpha c'' \cdot \mathbb{1}_{R>s} & rp'(v'' - c'') \end{bmatrix}. \tag{10}$$

The determinant of $J_{s,t}$ is given by $det(J_{s,z}) = \gamma r p'' v'' - c''(\gamma r p'' + \alpha^2 v'')$. Note that due to the assumptions on the convexity and concavity of the functions, this determinant is always positive. Taking into account that $det(J_{s,z})$ is positive, and the signs of the different derivatives, we obtain that the signs of the comparative statics for the case $t^* = 0$ are:

$$\text{sign}\begin{pmatrix} \frac{\partial s^*}{\partial \Lambda} & \frac{\partial s^*}{\partial \gamma} \\ \frac{\partial z^*}{\partial \Lambda} & \frac{\partial z^*}{\partial \gamma} \end{pmatrix} = \begin{pmatrix} +\mathbb{1}_{R>s} & - \\ -\mathbb{1}_{R>s} & + \end{pmatrix}.$$

For the case $t^* > 0$, we define function G using Equations 4, 6 and 7, and obtain the Jacobians with respect to s, t, z and Λ, γ, respectively:

$$J_{s,t,z} = \begin{bmatrix} v'' & v'' - \gamma\kappa'' & 0 \\ c'' & 0 & \alpha c'' - \frac{\gamma r p''}{\alpha} \\ 0 & \kappa'' & -\frac{rp''}{\alpha} \end{bmatrix}, \qquad J_{\Lambda,\gamma} = \begin{bmatrix} \mathbb{1}_{R>s+t} & -\kappa' \\ 0 & -\frac{rp'}{\alpha} \\ 0 & 0 \end{bmatrix}.$$

We use once more the Implicit Function Theorem to compute the comparative statics of s, t, z with respect to Λ and γ:

$$\begin{bmatrix} \frac{\partial s^*}{\partial \Lambda} & \frac{\partial s^*}{\partial \gamma} \\ \frac{\partial t^*}{\partial \Lambda} & \frac{\partial t^*}{\partial \gamma} \\ \frac{\partial z^*}{\partial \Lambda} & \frac{\partial z^*}{\partial \gamma} \end{bmatrix} = -J^{-1}_{s,t,z} \times J_{\Lambda,\gamma} =$$

$$= -\frac{1}{det(J_{s,t,z})} \begin{bmatrix} -\kappa''\left(\alpha c'' - \frac{\gamma r p''}{\alpha}\right) & \frac{rp''}{\alpha}(v'' - \gamma\kappa'') & (v'' - \gamma\kappa'')\left(\alpha c'' - \frac{\gamma r p''}{\alpha}\right) \\ \frac{rc''p''}{\alpha} & -\frac{rp''v''}{\alpha} & -v''\left(\alpha c'' - \frac{\gamma r p''}{\alpha}\right) \\ \kappa'' c'' & -\kappa'' v'' & -c''(v'' - \gamma\kappa'') \end{bmatrix} \times \begin{bmatrix} \mathbb{1}_{R>s+t} & -\kappa' \\ 0 & -\frac{rp'}{\alpha} \\ 0 & 0 \end{bmatrix} =$$

$$= -\frac{1}{det(J_{s,t,z})} \begin{bmatrix} -\kappa''\left(c'' - \frac{\gamma r p''}{\alpha}\right)\mathbb{1}_{R>s+t} & \kappa'\kappa''\left(\alpha c'' - \frac{\gamma r p''}{\alpha}\right) - p'p''\frac{r^2}{\alpha^2}(v'' - \gamma\kappa'') \\ \frac{c''p''r}{\alpha}\mathbb{1}_{R>s+t} & -\frac{r\kappa' c'' p''}{\alpha} + p'p''v''\frac{r^2}{\alpha^2} \\ \kappa'' c'' \mathbb{1}_{R>s+t} & -\kappa'\kappa'' c'' + \frac{rp'\kappa'' v''}{\alpha}. \end{bmatrix}$$

The determinant of $J_{s,t,z}$ is given by $det(J_{s,t,z}) = \frac{\gamma r \kappa'' p'' v''}{\alpha} - \frac{c''}{\alpha}(\gamma r \kappa'' p'' + (\alpha^2\kappa'' - rp'')v'')$. Note that due to the assumptions on the functions, this determinant is always positive. Note also that in the final matrix, all of the entries have an unambiguous sign except for the one that corresponds to $\frac{\partial s}{\partial \gamma}$ (first row, second column,), that will be negative if and only if $\kappa'\kappa''\left(\alpha c'' - \frac{\gamma r p''}{\alpha}\right) > p'p''\frac{r^2}{\alpha^2}(v'' - \gamma\kappa'')$. Hence, we have the following signs for the comparative statics:

$$\text{sign}\begin{pmatrix} \frac{\partial s^*}{\partial \Lambda} & \frac{\partial s^*}{\partial \gamma} \\ \frac{\partial t^*}{\partial \Lambda} & \frac{\partial t^*}{\partial \gamma} \\ \frac{\partial z^*}{\partial \Lambda} & \frac{\partial z^*}{\partial \gamma} \end{pmatrix} = \begin{pmatrix} +\mathbb{1}_{R>s+t} & \pm \\ +\mathbb{1}_{R>s+t} & - \\ -\mathbb{1}_{R>s+t} & + \end{pmatrix}$$

Therefore, from the signs of the comparative statics we have derived, and taking into account that the *loss aversion* channel becomes stronger as Λ increases, and the *behavioral spillovers* as γ decreases, this concludes the proof. □

Proof of Proposition 2. Note that the extensive margin of theft being weakly increasing is equivalent to showing that the extensive margin of theft is not strictly decreasing, i.e. that it is not possible for a participant to go from $t^* > 0$ to $t^* = 0$. We will start by proving the *behavioral spillovers* case, that becomes stronger as γ decreases. Note that as γ decreases, the cost of stealing $\gamma\kappa(t)$ also decreases. Therefore, if a participant chose $t^* > 0$ for a certain value of γ, *a fortiori* will chose $t^* > 0$ for a lower value of γ, because the cost of theft is decreased. Hence, no participant goes from $t^* > 0$ to $t^* = 0$ as γ decreases.

For the *loss aversion* case, the reasoning is quite similar. If a participant chose $t^* > 0$ for a certain value of Λ, then *a fortiori* will chose $t^* > 0$ when the value of Λ increases, as the penalty for having combined income $s^* + t^* < R$, that is given by $-\Lambda[R - s^* - t^*]_+$, has increased. Hence, no participant goes from $t^* > 0$ to $t^* = 0$ as Λ increases. □

Proof of Proposition 3. The proof for the first part is quite straightforward. Recall that s^*, t^* and z^* are the solutions to the participant's maximization problem when there is no constraint, and s^{**}, t^{**} and z^{**} are the solutions when there is the constraint $s + t \geq R$. Let $\tilde{u}(s,t,z) = v(s+t) + \gamma r p(z) - c(s+\alpha z) - \gamma\kappa(t)$, i.e. the utility function without loss aversion. We can then define $\Lambda^* = \frac{\tilde{u}(s^*,t^*,z^*) - \tilde{u}(s^{**},t^{**},z^{**})}{R - s^* - t^*}$, and from this definition we immediately have that for any $\Lambda > \Lambda^*$ we must have that $u(s^{**}, t^{**}, z^{**}) \geq u(s^*, t^*, z^*) = \tilde{u}(s^*, t^*, z^*) - \Lambda(R - s^* - t^*)$.

The proof for $\gamma \to 0$ is as follows. Note that from Equation 5 (from the first order conditions), and from the fact that $0 \leq z \leq 1$, and therefore $p'(z)$ is bounded, we have that as $\gamma \to 0$, $\gamma r p'(z) \to 0$, and therefore that $c'(s + \alpha z) \to 0$. But that means that $s + \alpha z \to 0$ (as $c'(0) = 0$) and $c(s + \alpha z)$ is increasing and convex, and since both s and z are non-negative, that means $s^* \to 0$ and $z^* \to 0$. Now, from Equation 3, we have that as $\gamma \to 0$, $v'(s + t) \to \Lambda \mathbb{1}_{R>s+t}$. But note that when $t > \Lambda$, $\mathbb{1}_{R>s+t} = 0$, and so we have that $v'(s + t) \to 0$, and therefore $t \to \infty$ (since v is increasing and concave), what means that the participants' utility also converges to $\lim_{t\to\infty} v(t)$, and therefore this is indeed the optimal solution, as we wanted to show. □

Proof of Proposition 4. Let s^*_R and t^*_R be the solutions for an agent i with cost functions (c_i, κ_i), in the *Reward* treatment; and s^*_P and t^*_P in the *Penalty* treatment. We will show first that there is more bunching in *Penalty* at R under the *loss aversion* channel. Let $(c_i, \kappa_i) \in \mathcal{C}^+_{Reward}(\delta, R)$, so for that participant, $s^*_R + t^*_R \in (R, R + \delta)$. Then, it must be the case that $s^*_R + t^*_R = s^*_P + t^*_P$, as when $s^*_R + t^*_R > R$, *loss aversion* is irrelevant, and the maximization problems in both treatments are identical. This proves $\mathcal{C}^+_{Reward}(\delta, R) \subset \mathcal{C}^+_{Penalty}(\delta, R)$. To show $\mathcal{C}^-_{Reward}(\delta, R) \not\subset \mathcal{C}^-_{Penalty}(\delta, R)$, consider a (c_i, κ_i) such that the participant's solution is $s^*_R + t^*_R = R - \epsilon$, for ϵ small enough, as defined below. Note that we can define $s^*(\Lambda)$ as a function of Λ, and from Proposition 1, we know that for $s^*_P + t^*_P < R$, $\frac{\partial s^*(\Lambda)}{\partial \Lambda}$ is positive. Therefore, we have that

$$(11) \qquad s^*_P - s^*_R = s^*(\Lambda) - s^*(0) = \int_0^\Lambda \frac{\partial s^*(\lambda)}{\partial \lambda} d\lambda > 0.$$

Hence, for $\epsilon < s^*_P - s^*_R$, and given that $t^*_P \geq t^*_R$, we have that $s^*_P + t^*_P > s^*_R + t^*_R + \epsilon \geq R$, and thus $\mathcal{C}^-_{Reward}(\delta, R) \not\subset \mathcal{C}^-_{Penalty}(\delta, R)$.

To show that under the *loss aversion* channel it is not true for any $x \neq R$ that there is more bunching in *Penalty* at x:

- if $x > R$, then we can choose δ small enough such that $x - \delta > R$. But then, if $(c_i, \kappa_i) \in \mathcal{C}^{+}_{Reward}(\delta, R)$, it means that $s^*_R + t^*_R \in (R, R+\delta)$. Hence, $\mathbb{1}[R > s+t] = 0$, and irrespective of the value of Λ, we have $\Lambda \mathbb{1}[R > s + t] = 0$: therefore, the solution in the *Penalty* treatment is identical, so $s^*_R + t^*_R = s^*_P + t^*_P$, and hence $\mathcal{C}^{-}_{Reward}(\delta, x) = \mathcal{C}^{-}_{Penalty}(\delta, x)$, and therefore it is not true that $\mathcal{C}^{-}_{Reward}(\delta, x) \not\subset \mathcal{C}^{-}_{Penalty}(\delta, x)$;
- if $x < R$, following an argument similar to that we used in Equation 11, we can find (c_i, κ_i) such that $s^*_R + t^*_R = x + \delta - \epsilon$, for $\epsilon > 0$ small enough, such that $s^*_P + t^*_P \geq x + \delta$, and therefore it is not true that $\mathcal{C}^{+}_{Reward}(\delta, x) \subset \mathcal{C}^{+}_{Penalty}(\delta, x)$.

The argument to show that under the *behavioral spillovers* channel there is no point at which there is more bunching in *Penalty*, also follows a similar reasoning as we used in Equation 11 (recall that as the *behavioral spillovers* channel becomes activated, γ decreases). For any x:

- if $s^* + t^*$ is decreasing in γ, we can find (c_i, κ_i) such that $s^*_R + t^*_R = x + \delta - \epsilon$, for $\epsilon > 0$ small enough, such that $s^*_P + t^*_P \geq x + \delta$, and therefore it is not true that $\mathcal{C}^{+}_{Reward}(\delta, x) \subset \mathcal{C}^{+}_{Penalty}(\delta, x)$;
- if $s^* + t^*$ is increasing in γ, we can find (c_i, κ_i) such that $s^*_R + t^*_R = x + \epsilon$, for $\epsilon > 0$ small enough, such that $s^*_P + t^*_P \leq x$, and once again it is not true that $\mathcal{C}^{+}_{Reward}(\delta, x) \subset \mathcal{C}^{+}_{Penalty}(\delta, x)$.

□

Figures and tables for additional robustness analysis

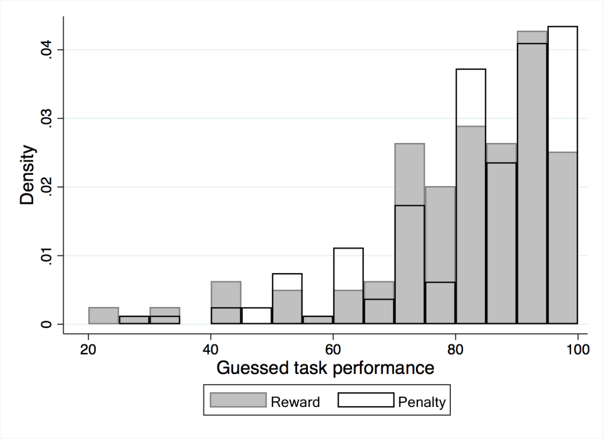


Figure 6: Distribution of self-reported guessed performance, by treatment

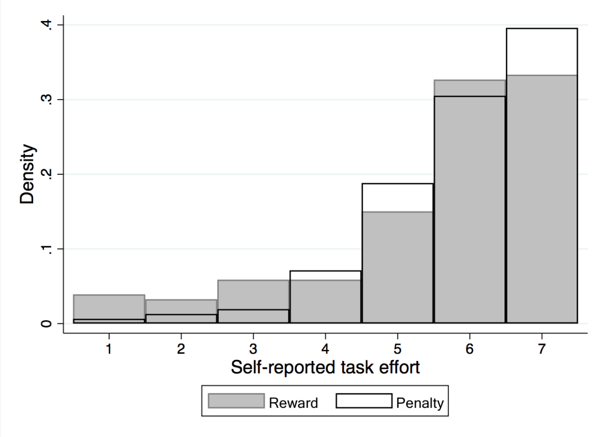
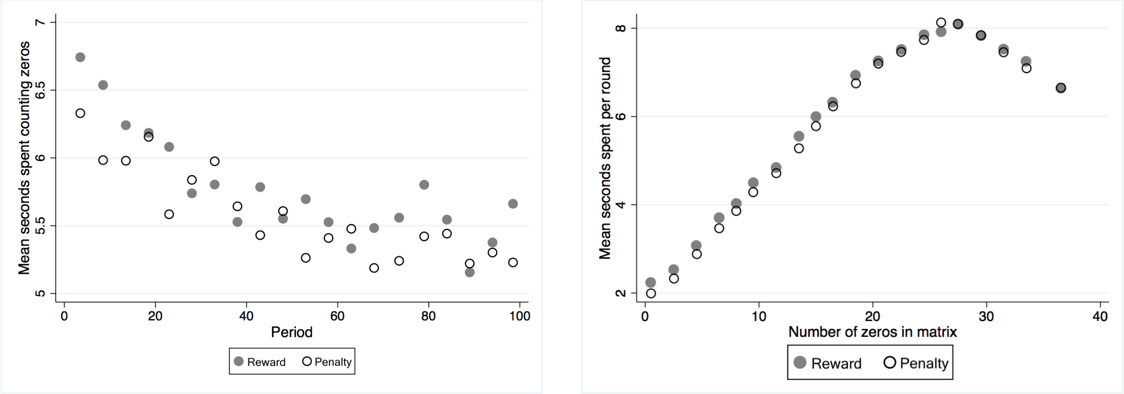


Figure 7: Distribution of self-reported effort provision, by treatment



(a) Over time

(b) By task difficulty

Figure 8: Seconds spent on calculation task per round

Table 3: Balance test

	Reward	*Penalty*	Difference
Male	0 .591	0.602	-0.011
	(0.039)	(0.039)	(0.055)
Age	21.765	22.039	-0.274
	(0.212)	(0.311)	(0.377)
Year of study	3.253	3.218	0.036
	(0 .125)	(0 .124)	(0.176)
Econ student	0.686	0.702	-0.016
	(0.037)	(0 .0362)	(0.052)
N	159	161	

Comparison of means using t-tests with unequal variances; * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Age and year of study values are missing for 13 and 1 student(s), respectively.

Table 4: Number of items stolen, by category and treatment

	Reward	*Penalty*
Pencil	8	10
Eraser	8	11
Sharpener	14	17
Yellow marker	11	13
Fine liner red	8	23
Fine liner blue	16	26
Post-it note	5	5
Pen	11	9
Correction roller	2	6
Total	**83**	**120**

References

Abeler, J., Falk, A., Goette, L., and Huffman, D. (2011). Reference points and effort provision. *American Economic Review*, 101(2):470–492.

Allen, E. J., Dechow, P. M., Pope, D. G., and Wu, G. (2017). Reference-dependent preferences: Evidence from marathon runners. *Management Science*, 63(6):1657–1672.

Armantier, O. and Boly, A. (2015). Framing of incentives and effort provision. *International Economic Review*, 56(3):917–938.

Association of Certified Fraud Examiners, T. (2016). Report to the Nations on Occupational Fraud and Abuse: 2016 Global Fraud Study. Technical report.

Barberis, N. C. (2013). Thirty Years of Prospect Theory in Economics: A Review and Assessment. *Journal of Economic Perspectives*, 27(1):173–196.

Barbos, A. (2010). Context effects: A representation of choices from categories. *Journal of Economic Theory*, 145(3):1224–1243.

Belot, M. and Schröder, M. (2013). Sloppy work, lies and theft: A novel experimental design to study counterproductive behaviour. *Journal of Economic Behavior and Organization*, 93:233–238.

Belot, M. and Schröder, M. (2016). The Spillover Effects of Monitoring: A Field Experiment. *Management Science*, 62(1):37–45.

Bowles, S. (1998). Endogenous preferences: The cultural consequences of markets and other economic institutions. *Journal of Economic Literature*, 36(1):75–111.

Bradler, C. and Neckermann, S. (2016). The Magic of the Personal Touch: Field Experimental Evidence on Money and Appreciation as Gifts.

Breza, E., Kaur, S., and Shamdasani, Y. (2018). The morale effects of pay inequality. *Quarterly Journal of Economics*, 133(2):611–663.

Brooks, R. R. W., Stremitzer, A., and Tontrup, S. (2012). Framing Contracts: Why Loss Framing Increases Effort. *Journal of Institutional and Theoretical Economics (JITE)*, 168(1):62–82.

Brownback, A. and Sadoff, S. (2019). Improving College Instruction Through Incentives. *Journal of Political Economy.*

Bulte, E., List, J. A., and Van Soest, D. (2019). Toward an Understanding of the Welfare Effects of Nudges: Evidence from a Field Experiment in Uganda. Technical Report 26286, National Bureau of Economic Research.

Buser, T. and Dreber, A. (2016). The Flipside of Comparative Payment Schemes. *Management Science*, 62(9):2626–2638.

Cameron, J. S. and Miller, D. T. (2009). Ethical standards in gain versus loss frames. In De Cremer, D., editor, *Psychological perspectives on ethical behavior*, pages 91–106. Information Age Publishing.

Chetty, R., Friedman, J. N., Olsen, T., and Pistaferri, L. (2011). Adjustment costs, firm-responses, and micro vs. macro laborsupply elasticities: Evidence fromdanish tax records. *Quarterly Journal of Economics*, 126(2):749–804.

Dana, J., Weber, R. A., and Kuang, J. X. (2007). Exploiting moral wiggle room: Experiments demonstrating an illusory preference for fairness. In *Economic Theory*, volume 33, pages 67–80. Springer-Verlag.

Danilov, A. and Vogelsang, T. (2016). Time for helping. *Journal of the Economic Science Association*, 2(1):36–47.

De Giorgi, E. G. and Post, T. (2011). Loss aversion with a state-dependent reference point. *Management Science*, 57(6):1094–1110.

De Quidt, J. (2017). Your Loss Is My Gain: A Recruitment Experiment with Framed Incentives. *Journal of the European Economic Association*, 51(5):351–365.

De Quidt, J., Fallucchi, F., Kölle, F., Nosenzo, D., and Quercia, S. (2017). Bonus versus penalty: How robust are the effects of contract framing? *Journal of the Economic Science Association*, pages 1–9.

DellaVigna, S. and Pope, D. (2017). What Motivates Effort? Evidence and Expert Forecasts — The Review of Economic Studies. *The Review of Economic Studies.*

Dur, R. (2009). Gift exchange in the workplace: Money or attention? *Journal of the European Economic Association*, 7(2-3):550–60.

Englmaier, F., Grimm, S., Schindler, D., and Schudy, S. (2018). The Effect of Incentives in Non-Routine Analytical Teams Tasks-Evidence from a Field Experiment. Technical Report 6903.

Fehr, E. and Schmidt, K. M. (2006). Chapter 8 The Economics of Fairness, Reciprocity and Altruism - Experimental Evidence and New Theories.

Fischbacher, U. (2007). z-Tree: Zurich toolbox for ready-made economic experiments. *Experimental Economics*, 10(2):171–178.

Fischbacher, U. and Föllmi-Heusi, F. (2013). Lies in disguise-an experimental study on cheating. *Journal of the European Economic Association*, 11(3):525–547.

Fryer, R., Levitt, S., List, J., and Sadoff, S. (2012). Enhancing the Efficacy of Teacher Incentives through Loss Aversion: A Field Experiment.

Gall, T., Hu, X., and Vlassopolous, M. (2016). Dynamic Incentive Effects of Team Formation: Experimental Evidence.

Gilboa, I., Minardi, S., and Wang, F. (2020). Consumption of Values. *HEC Paris Research Paper No. ECO/SCD-2020-1406*.

Gneezy, U. and Imas, A. (2014). Materazzi effect and the strategic use of anger in competitive interactions. *Proceedings of the National Academy of Sciences of the United States of America*, 111(4):1334–7.

Gneezy, U., Kajackaite, A., and Sobel, J. (2018). Lying aversion and the size of the lie. *American Economic Review*, 108(2):419–53.

Gneezy, U., Meier, S., and Rey-Biel, P. (2011). When and Why Incentives (Don't) Work to Modify Behavior. *Journal of Economic Perspectives*, 25(4):191–210.

Gneezy, U. and Rustichini, A. (2000). Pay Enough or Don't Pay at All. *The Quarterly Journal of Economics*, 115(3):791–810.

Goette, L., Huffman, D., Meier, S., and Sutter, M. (2012). Competition Between Organizational Groups: Its Impact on Altruistic and Antisocial Motivations. *Management Science*, 58(5):948–960.

Gravert, C. (2013). How luck and performance affect stealing. *Journal of Economic Behavior and Organization*, 93:301–304.

Grolleau, G., Kocher, M. G., and Sutan, A. (2016). Cheating and Loss Aversion: Do People Cheat More to Avoid a Loss? *Management Science*, 62(12):3428–3438.

Guney, B., Richter, M., and Tsur, M. (2018). Aspiration-based choice. *Journal of Economic Theory*, 176:935–956.

Hannan, R. L., Hoffman, V. B., and Moser, D. V. (2005). Bonus versus Penalty: Does Contract Frame Affect Employee Effort? In *Experimental Business Research*, volume II, pages 151–169. Springer-Verlag, Berlin/Heidelberg.

Harbring, C. and Irlenbusch, B. (2011). Sabotage in Tournaments: Evidence from a Laboratory Experiment. *Management Science*, 57(4):611–627.

Hermann, D. and Mußhoff, O. (2019). I might be a liar, but I am not a thief: An experimental distinction between the moral costs of lying and stealing. *Journal of Economic Behavior & Organization*, 163:135–139.

Holmstrom, B. and Milgrom, P. (1991). Multitask Principal-Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design.

Hong, F., Hossain, T., and List, J. A. (2015). Framing manipulations in contests: A natural field experiment. *Journal of Economic Behavior and Organization*, 118:372–382.

Hossain, T. and List, J. A. (2012). The Behavioralist Visits the Factory: Increasing Productivity Using Simple Framing Manipulations. *Management Science*, 58(12):2151–2167.

Hsee, C. K., Yu, F., Zhang, J., and Zhang, Y. (2003). Medium Maximization. *Journal of Consumer Research*, 30(1):1–14.

Imas, A. (2016). The realization effect: Risk-taking after realized versus paper losses. *American Economic Review*, 106(8):2086–2109.

Imas, A., Sadoff, S., and Samek, A. (2017). Do People Anticipate Loss Aversion? *Management Science*, 63(5):1271–1284.

Kahneman, D. and Tversky, A. (1979). Prospect Theory: An Analysis of Decision under Risk. *Econometrica*, 47(2):263.

Kajackaite, A. and Gneezy, U. (2017). Incentives and cheating. *Games and Economic Behavior*, 102:433–444.

Kern, M. C. and Chugh, D. (2009). Bounded Ethicality: The Perils of Loss Framing. *Psychological Science*, 20(3):378–384.

Kleven, H. J. (2016). Bunching. *Annual Review of Economics*, 8:435–464.

Koszegi, B. (2006). Emotional Agency. *The Quarterly Journal of Economics*, 121(1):121–155.

Levitt, S. D., List, J. A., Neckermann, S., and Sadoff, S. (2016). The behavioralist goes to school: Leveraging behavioral economics to improve educational performance. *American Economic Journal: Economic Policy*, 8(4):183–219.

Loewenstein, G. (2000). Emotions in Economic Theory and Economic Behavior.

Maltz, A. (2020). Exogenous Endowment-Endogenous Reference Point∗. *Economic Journal*, 130(625):160–182.

Masatlioglu, Y. and Ok, E. A. (2014). A canonical model of choice with initial endowments. *Review of Economic Studies*, 81(2):851–883.

Mazar, N., Amir, O., and Ariely, D. (2008). The Dishonesty of Honest People: A Theory of Self-Concept Maintenance. *Journal of Marketing Research*, 45(6):633–644.

Neckermann, S., Cueni, R., and Frey, B. S. (2014). Awards at work. *Labour Economics*, 31:205–217.

Ockenfels, A., Sliwka, D., and Werner, P. (2015). Bonus payments and reference point violations. *Management Science*, 61(7):1496–1513.

Ok, E. A., Ortoleva, P., and Riella, G. (2015). Revealed (p) reference theory. *American Economic Review*, 105(1):299–321.

Ortoleva, P. (2010). Status quo bias, multiple priors and uncertainty aversion. *Games and Economic Behavior*, 69(2):411–424.

Pettit, N. C., Doyle, S. P., Lount, R. B., and To, C. (2016). Cheating to get ahead or to avoid falling behind? The effect of potential negative versus positive status change on unethical behavior. *Organizational Behavior and Human Decision Processes*, 137:172–183.

Pierce, L., Rees-Jones, A., and Blank, C. (2020). The Negative Consequences of Loss-Framed Performance Incentives. Working Paper 26619, National Bureau of Economic Research.

Pierce, L., Snow, D. C., and McAfee, A. (2015). Cleaning house: The impact of information technology monitoring on employee theft and productivity. *Management Science*, 61(10):2299–2319.

Podsakoff, P. M., MacKenzie, S. B., Paine, J. B., and Bachrach, D. G. (2000). Organizational Citizenship Behaviors: A Critical Review of the Theoretical and Empirical Literature and Suggestions for Future Research. *Journal of Management*, 26(3):513–563.

Rabin, M. (1994). Cognitive dissonance and social change. *Journal of Economic Behavior and Organization*, 23(2):177–194.

Riella, G. and Teper, R. (2014). Probabilistic dominance and status quo bias. *Games and Economic Behavior*, 87:288–304.

Rubin, J., Samek, A., and Sheremeta, R. M. (2018). Loss aversion and the quantity–quality tradeoff. *Experimental Economics*, 21(2):292–315.

Ruggeri, K., Alí, S., Berge, M. L., Bertoldo, G., Bjørndal, L. D., Cortijos-Bernabeu, A., Davison, C., Demić, E., Esteban-Serna, C., Friedemann, M., Gibson, S. P., Jarke, H., Karakasheva, R., Khorrami, P. R., Kveder, J., Andersen, T. L., Lofthus, I. S., McGill, L., Nieto, A. E., Pérez, J., Quail, S. K., Rutherford, C., Tavera, F. L., Tomat, N., Reyn, C. V., Većkalov, B., Wang, K., Yosifova, A., Papa, F., Rubaltelli, E., van der Linden, S., and Folke, T. (2020). Replicating patterns of prospect theory for decision under risk. *Nature Human Behaviour*, pages 1–12.

Schindler, S. and Pfattheicher, S. (2017). The frame of the game: Loss-framing increases dishonest behavior. *Journal of Experimental Social Psychology*, 69:172–177.

Schurr, A. and Ritov, I. (2016). Winning a competition predicts dishonest behavior. *Proceedings of the National Academy of Sciences of the United States of America*, 113(7):1754–1759.

Shalvi, S. (2012). Dishonestly increasing the likelihood of winning. *Judgement and Decision Making*, 7(3):292–303.

Shalvi, S., Handgraaf, M. J. J., and De Dreu, C. K. (2011). Ethical Manoeuvring: Why People Avoid Both Major and Minor Lies. *British Journal of Management*, 22(s1):S16–S27.

Thaler, R. H. and Johnson, E. J. (1990). Gambling with the House Money and Trying to Break Even: The Effects of Prior Outcomes on Risky Choice. *Management Science*, 36(6):643–660.

Tversky, A. and Kahneman, D. (1991). Loss aversion in riskless choice: A reference-dependent model. *The Quarterly Journal of Economics*, 106(4):1039–1061.

Young, A. (2019). Channeling Fisher: Randomization tests and the statistical insignificance of seemingly significant experimental results. *The Quarterly Journal of Economics*, 134(2):557–98.

Online Appendix: “Loss-framed incentives and employee (mis-)behavior”

This document supplements the main text of the article “Loss-framed incentives and employee (mis-)behavior”.

Section A provides further details on the experimental procedure. Section A.1 provides photographs detailing the experimental procedure. Section A.2 provides the experimental instructions. Section A.3 contains the script for the obligatory questionnaire. Section A.4 contains the voluntary survey.

Section B offers additional tables and figures for the secondary analysis of participant experience we performed in Section 5.5, where we used participants’ answers in the obligatory questionnaire to explore which of the two channels drive behavior in response to loss-framed incentives.

Appendix A Experimental Procedure

A.1 Set-up of the experiment

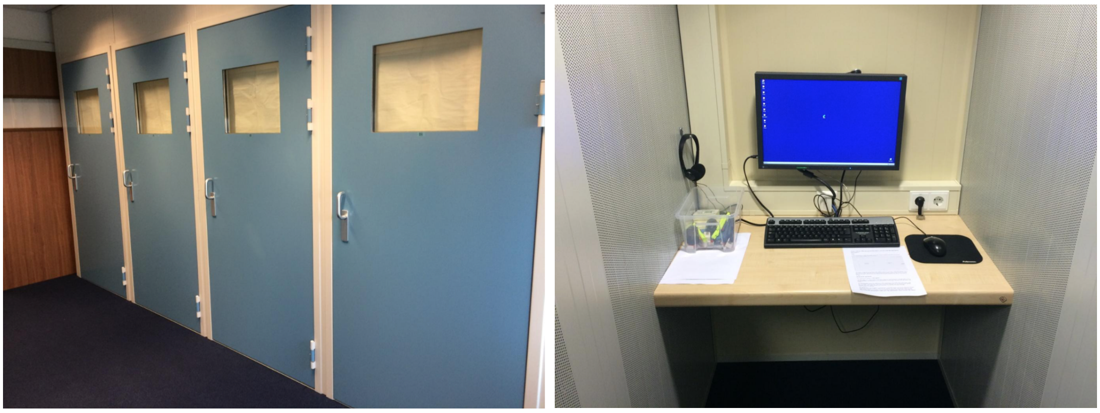


Figure A1: The cubicles in the experimental laboratory

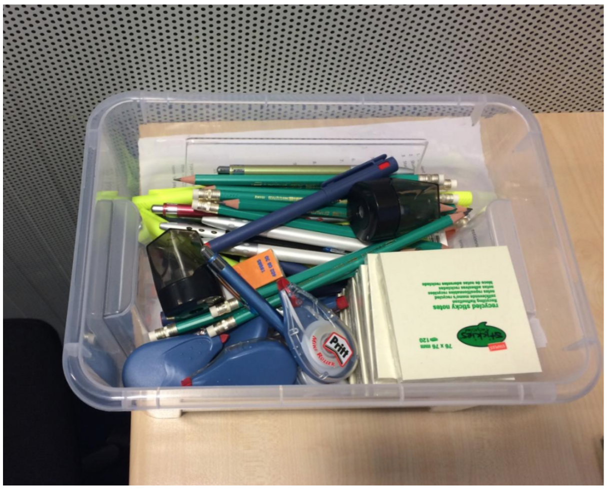
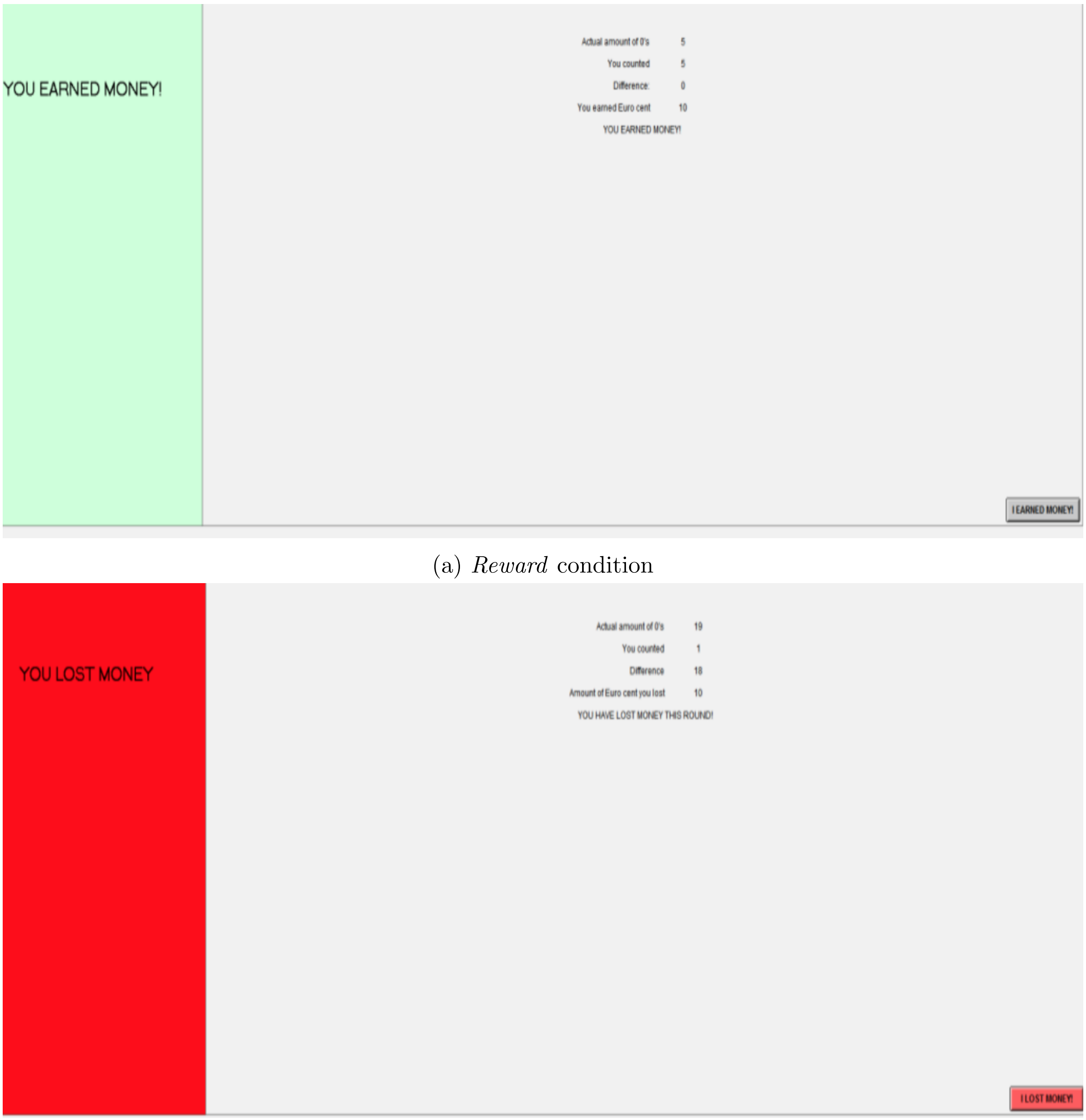


Figure A2: The box of office supplies placed in each cubicle



YOU EARNED MONEY!

Actual amount of 0's 5
You counted 5
Difference: 0
You earned Euro cent 10
YOU EARNED MONEY!

(a) *Reward* condition

YOU LOST MONEY

Actual amount of 0's 19
You counted 1
Difference 18
Amount of Euro cent you lost 10
YOU HAVE LOST MONEY THIS ROUND!

(b) *Penalty* condition

Figure A3: Screenshots of the feedback page in the task

We include here the instructions that were provided to the subjects. The instructions were identical in both treatments, except for the payment structure part, where we have written the instructions for the Reward treatment in green, and those for the Penalty treatment in red (text in between << >> is for context and was not included in the subject's instructions).

A.2 Experimental Instructions

Please read the following instructions before you start with the experiment!

Set-up of the experiment

The experiment consists of **three parts.**

The **first part** are 5 trial rounds of the task so that you can familiarize yourself with it. There are no monetary consequences to your performance in this part. This part will take about one minute.

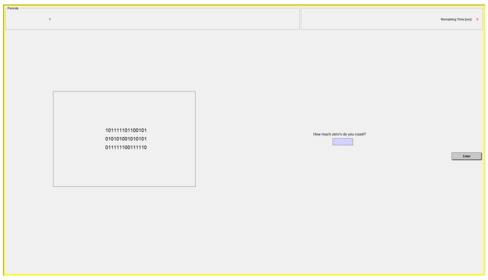
The **second part** will be the main part of the experiment. You will work on the task for 100 rounds. Your final payment in the experiment depends on your performance during this part. We will explain the payment structure below. This part will take approximately 25 minutes.

The **third part** is an obligatory questionnaire. You find it at the top of your desk under the container with the pencils. It is one page long and will take about 1 or 2 minutes to fill in. There are also pencils and other material provided on your desk. Please check now whether you see both the questionnaire and a box with material. **Please bring this questionnaire with you to the experimenter when you leave the cubicle.**

Finally, you could help with a different survey. Participation in this additional survey is voluntary and there will be no reward or punishment for it. If you are willing to help, feel free to fill in the survey that you find under the questionnaire.

Task

In this experiment you will work on the matrix task. When working on the task, you will get to see a screen that is going to be similar to this:



101111101100101
010101001010101
011111100111110

How much zero's do you count?

The object on the left shows rows with 0's and 1's. Your task is to enter the amount of 0's into the box on the right side of the screen and you **have to press the "Enter" button on the screen.** Only then will your answer be registered in the system. You will have 10 seconds to do this.

Right after that you will see a screen which shows you the correct answer, your answer, and the payoff consequences of this. Please press the button at the bottom of the screen to proceed. Otherwise, the program will continue automatically after 10 seconds.

Payment structure

<< *Reward treatment* >>

Your payment depends on your performance in the second part of the experiment. For every task that you do correctly, you **earn** 10 cents. At the end of the experiment, you will receive the sum of earnings from all your correct answers.

For example, if you solve 50 matrices correctly, you will earn 5 Euros, which you will receive in cash at the end of the experiment.

<< *Penalty treatment* >>

For participating in this experiment you have already received 10.00€. These are yours. However, for every task that you do not do correctly, you incur a loss of 0.10€. At the end of the experiment, the sum of all your wrong answers will be **deducted** and you will have to **pay the experimenter back** from the money that you already received.

For example, if you solve 10 matrices incorrectly, you have to pay 1.00€ in cash to the experimenter. Cash change is available.

The experimenter will remain in the experimenter room throughout the entire experiment. If you have a question, please go ask him there.

If you use the computer in an improper way you will be excluded from the experiment and from any payment.

Please close your door. The experiment will automatically start in a few seconds.

A.3 Obligatory questionnaire

Obligatory Questionnaire

(pencils and other material are provided on desk)

1. Student name: ....

2. Student number: ....

3. Student age: ....

4. What year of study are you in?

Bachelor 1 Bachelor 2 Bachelor 3 Pre-Master Master Post-Master

5. What is your field of study?

Economics Business Psychology Law Other

6. What is your gender?

male female

7. We will invite some people back for another round of the same experiment within the next few weeks. Do you want to participate again?

yes; e-mail address: .... no

8. On a scale of 1 to 7, how happy are you now?

(1: not happy at all; 7: very happy) ....

9. On a scale of 1 to 7, how much fun was part two of the experiment?

(1: no fun at all; 7: a lot of fun) ....

10. Out of the 100 matrices you were presented with, how many counts do you think you got right in total?

....

11. On a scale of 1 to 7, how adequate/fair do you perceive the payment?

(1: completely inadequate/unfair; 7: completely adequate/fair) ....

12. On a scale of 1 to 7, how hard did you work on the task?

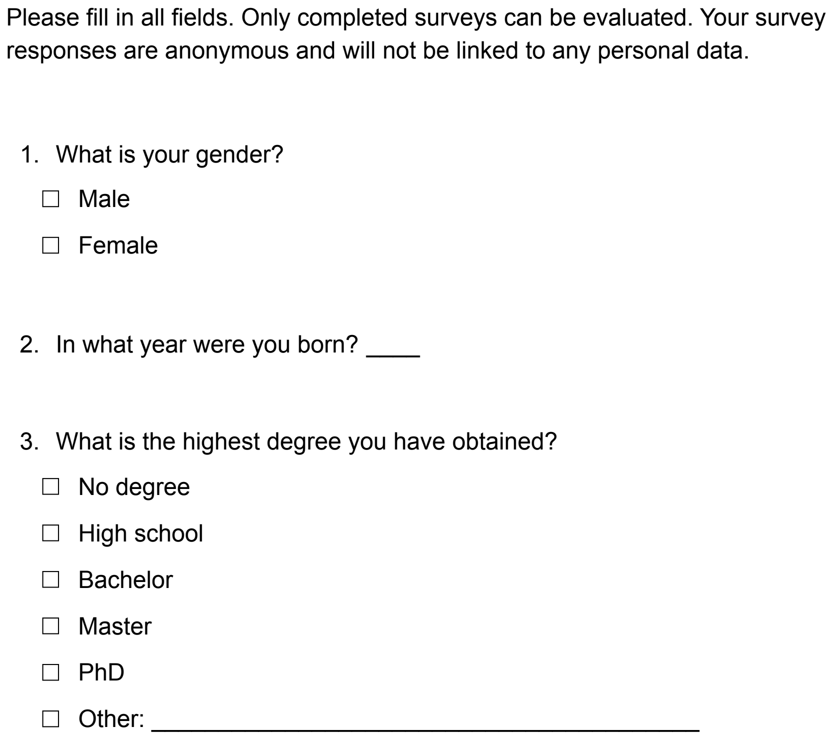
(1: not hard at all; 7: as hard as i could) ....

13. Would you suggest to your friends to participate in this experiment?

yes no

You could help us with another research project by filling in the survey that you find on your desk. It should take approximately 5 minutes. Otherwise, please proceed to experimenter room for payment.

A.4 Voluntary survey



Please fill in all fields. Only completed surveys can be evaluated. Your survey responses are anonymous and will not be linked to any personal data.

1. What is your gender?

   □ Male

   □ Female

2. In what year were you born? ____

3. What is the highest degree you have obtained?

   □ No degree

   □ High school

   □ Bachelor

   □ Master

   □ PhD

   □ Other: ____

The following questions are about your general opinion about motivation factors at work. For the answer, t is not necessary that you are currently working!

4a. **How important are the following factors in a job to you?**

Not important	1	2	3	4	5	6	7	ıportant
Dynamic environment	□	□	□	□	□	□	□	
High wage	□	□	□	□	□	□	□	
Good work relationship with colleagues and superiors	□	□	□	□	□	□	□	
Small gestures (eg, small gifts for a birthday or Christmas)	□	□	□	□	□	□	□	
Possibility to get additional leave-time	□	□	□	□	□	□	□	
Acces to unlimited trainings	□	□	□	□	□	□	□	
Appreciation and recognition from superior	□	□	□	□	□	□	□	
Doing something good for the world	□	□	□	□	□	□	□	
Flexible work hours	□	□	□	□	□	□	□	
Opportunity to work from home	□	□	□	□	□	□	□	

4b. Companies are looking for committed and motivated employees. Imagine you have a job you greatly enjoy. In what way will the following factors affect your job performance?

	Performance decreases			No effect		Performance increases	
	1	2	3	4	5	6	7
No monitoring	□	□	□	□	□	□	□
Flexible work hours	□	□	□	□	□	□	□
Possibility to work from home	□	□	□	□	□	□	□

4c. Imagine you have a job you do not enjoy. In what way will the following factors affect your job performance?

	Performance decreases			No effect		Performance increases	
	1	2	3	4	5	6	7
No monitoring	□	□	□	□	□	□	□
Flexible work hours	□	□	□	□	□	□	□
Possibility to work from home	□	□	□	□	□	□	□

4d. What do you think are other factors that are generally critical to employee motivation? (the more detailed your reply, the more helpful)

____

____

____

____

5. Imagine you have a job in which you can work from home.
Which statement best describes you?

□ I would (almost) always work from home

□ I would still work partly at the company

□ I would still work mainly at the company

⟶ If you would still go the company to work, what are your main reasons?

	I fully disagree	1	2	3	4	5	6	7	ˈ fully ıgree
I can focus more on the job in the company.		□	□	□	□	□	□	□	
I like to have personal contact with my colleagues.		□	□	□	□	□	□	□	
I find it hard to motivate myself to work at home.		□	□	□	□	□	□	□	
I let myself be distracted very easily while working at home.		□	□	□	□	□	□	□	

Other reasons: ____________________________________

6. Imagine you had a job as an employee, in which you needed to work fixed, predetermined hours (as in most professions). Now your employer allows you (to some extent) to freely decide when you want to work (time of day and day of week) as long as your total working hours remain the same. Which statement best describes your reaction:

□ I would probably use the freedom to adjust the work hours to my needs.

□ I would probably keep working on the fixed, predetermined schedule.

⟶ In the latter case, what is your motivation?

□ I like routine and stucture.

□ I am not good at time management.

Other reasons: ____________________________________

Please only fill in this page, if you are currently employed or have been employed at some point in the past; if you have never worked, you are done filling in this survey!

7a. What kind of work do you do now or have you been practicing mainly in the past?

Please provide the exact title of your occupation, e.g. 'Salesman' instead of 'employee' or 'police officer' instead of 'public sector'. If you are following a trainee- or apprenticeship, please enter that.

□

7b. Does your job allow flexible working hours, such as "Flextime"?

□ Yes, I have flexible working hours.

□ No, it would be possible in my profession, but my employer does not offer it.

□ No, it would not be possible in my profession.

7c. Do you have the possibility to work from home?

□ Yes

□ No, it is possible, but my employer wants me to be at the company during work hours

□ No, it is not possible (eg because I have to be at the production site or at the customer).

If 'Yes', to what extent? □ days per week

If 'No', would you like to work more from home?

□ Yes □ days more per week

□ No

7d. How satisfied are you with your current job?

Not satisfied at all	1	2	3	4	5	6	7	Very satisfied
	□	□	□	□	□	□	□	

7e. How much do you enjoy your current job?

No enjoyment at all	1	2	3	4	5	6	7	enjoy it a lot
	□	□	□	□	□	□	□	

Appendix B Additional tables and figures for the secondary analysis of participant experience

In Section 5.5, we analyze the answers to all five questions in the obligatory questionnaire (included in Section A.3 of this Online Appendix) that pertain to participants' experience in the experiment: how happy they feel, how fun the task was, how fair the compensation was, whether they would be willing to return for another round of the same experiment and whether they would be willing to recommend the experiment to their friends. Tables B1 and B2 show the correlations between the different measures of participant experience.

Table B1: Pairwise correlations in the *Reward* treatment

	Happy	Fun	Fair	Friends	Task performance	Stole	Survey complete	Return
Happy	1.000							
Fun	0.474***	1.000						
	(0.000)							
Fair	0.390***	0.354***	1.000					
	(0.000)	(0.000)						
Return	0.255**	0.215	0.301***	0.226	0.120	-0.084	0.117	1.000
	(0.034)	(0.183)	(0.003)	(0.117)	(1.000)	(1.000)	(1.000)	
Friends	0.362***	0.417***	0.419***	1.000				
	(0.000)	(0.000)	(0.000)					
Task performance	0.138	0.056	0.230*	0.063	1.000			
	(1.000)	(1.000)	(0.100)	(1.000)				
Stole	-0.042	0.015	-0.101	-0.037	-0.133	1.000		
	(1.000)	(1.000)	(1.000)	(1.000)	(1.000)			
Survey complete	-0.007	0.222	-0.019	0.124	0.035	0.104	1.000	
	(1.000)	(0.151)	(1.000)	(1.000)	(1.000)	(1.000)		

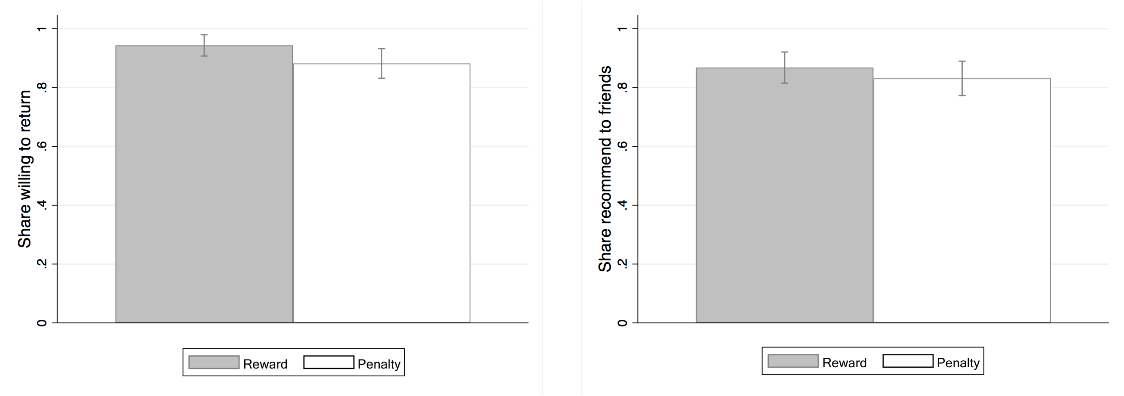
p-values in parentheses, using Bonferroni-adjusted significance levels; * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table B2: Pairwise correlations in the *Penalty* treatment

	Happy	Fun	Fair	Friends	Task performance	Stole	Survey complete	Return
Happy	1.000							
Fun	0.394***	1.000						
	(0.000)							
Fair	0.259**	0.198	1.000					
	(0.027)	(0.328)						
Return	0.257**	0.308***	0.176	0.350***	0.185	0.022	0.147	1.000
	(0.029)	(0.002)	(0.707)	(0.000)	(0.525)	(1.000)	(1.000)	
Friends	0.303***	0.483***	0.228	1.000				
	(0.003)	(0.000)	(0.104)					
Task performance	0.269**	0.150	0.348***	0.152	1.000			
	(0.016)	(1.000)	(0.000)	(1.000)				
Stole	0.064	-0.064	0.036	0.010	-0.002	1.000		
	(1.000)	(1.000)	(1.000)	(1.000)	(1.000)			
Survey complete	0.023	0.129	-0.005	0.140	-0.064	-0.009	1.000	
	(1.000)	(1.000)	(1.000)	(1.000)	(1.000)	(1.000)		

p-values in parentheses, using Bonferroni-adjusted significance levels; * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Because the answers to the five questions are highly correlated, as showed in Tables B1 and B2, we summarize them in a "participant satisfaction" index that corresponds to the first principal component of the five variables, rather than performing five individual comparisons. This index ranges from -6.2 to 2.4, with a mean of zero (by construction) and standard deviation of 1.5. According to this index, there is no meaningful difference between participants' experience between the two treatments: mean participants' satisfaction is slightly lower in the *Penalty* than in the *Reward* treatment, but the difference is small in size (0.107 SD) and not statistically significantly different from zero (p-value 0.339, randomization inference-based p-value 0.343). Analyzing the answers to each of the five survey questions separately confirms this conclusion. Figures B1a, B1b and B2 compare the answers to each question separately by treatment, and confirm that the differences in these self-reported outcome measures between the treatments are negligible in size.



(a) Return for another session

(b) Recommend experiment to friends

Figure B1: Share willing to return/recommend by treatment, with 95% CIs

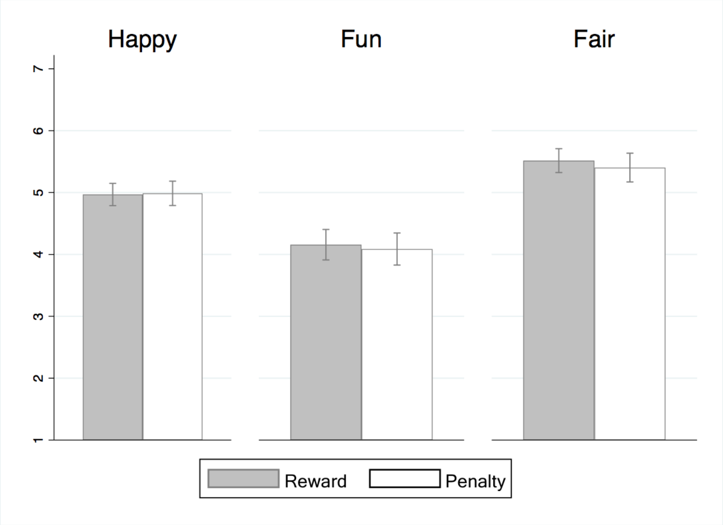


Figure B2: Mean ratings for the experiment by treatment, with 95% CIs

Finally, Figure B3 compares the distribution of the participant satisfaction index between treatments, and shows no meaningful difference in participant experience in the *Penalty* vs. the *Reward* condition.

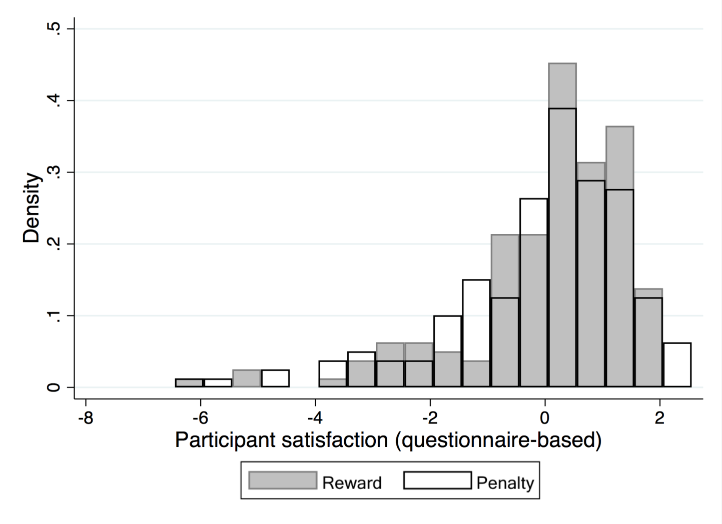


Figure B3: Distribution of participant satisfaction, by treatment