Downloaded from informs.org by [154.59.124.38] on 09 June 2023, at 05:30 . For personal use only, all rights reserved.

informs
https://pubsonline.informs.org/journal/mnsc

MANAGEMENT SCIENCE
Vol. 68, No. 10, October 2022, pp. 7518–7537
ISSN 0025-1909 (print), ISSN 1526-5501 (online)

Loss-Framed Incentives and Employee (Mis-)Behavior

Eszter Czibor,[a,*] Danny Hsu,[b] David Jimenez-Gomez,[c] Susanne Neckermann,[d] Burcu Subasi[e]

[a] Consortium on Compensating Income Variation, Department of Economics, University of Iceland, 102 Reykjavík, Iceland; [b] Erasmus School of Economics, 3062 PA Rotterdam, Netherlands; [c] Department of Economics, University of Alicante, 03690 San Vicente del Raspeig, Alicante, Spain; [d] University of Chicago, Chicago, Illinois 60637; [e] School of Financial and Economic Management, Hanze University of Applied Sciences, 9747 AS Groningen, Netherlands
*Corresponding author

Contact: czibore@gmail.com, https://orcid.org/0000-0002-3182-6564 (EC); dannyhsu1986@gmail.com (DH); davidjimenezgomez@ua.es, https://orcid.org/0000-0001-6664-7913 (DJ-G); sneckermann@uchicago.edu (SN); b.subasi@rug.nl (BS)

Received: November 8, 2018
Revised: June 22, 2020; July 23, 2021; October 20, 2021
Accepted: October 29, 2021
Published Online in Articles in Advance: February 17, 2022

https://doi.org/10.1287/mnsc.2021.4280

Copyright: © 2022 INFORMS

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 framed as either a reward (gain) or 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 it 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.

History: Accepted by Axel Ockenfels, behavioral economics and decision analysis.

Funding: David Jimenez-Gomez is grateful for the funding from the Ministry of Science and Innovation and the European Regional Development Fund [Grant PID2019-107081GB-I00], from the Ministry of Science and Innovation through the project "Nudging applied to the improvement of regulation" [Grant RED2018-102761-T], and from the Valencian Community through the Prometeo program [Grant PROMETEO/2021/073].

Supplemental Material: The online appendix and data are available at https://doi.org/10.1287/mnsc.2021.4280.

Keywords: loss-framed incentives • multitasking • incentive design • stealing

1. Introduction

Loss-framed incentives have received a lot of attention in behavioral economics. In contrast to traditional gain-framed incentive schemes in which employees receive a reward upon achieving a prespecified target, loss-framed incentive schemes entail an up-front payment to the employees, which they lose if they fail to reach the target.[1] Several papers find that loss-framed incentives lead to greater effort provision than their gain-framed equivalents (e.g., Hannan et al. 2005, Fryer et al. 2012, Hossain and List 2012, Armantier and Boly 2015, Levitt et al. 2016, Bulte et al. 2019). 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 in which employees have more than one strategy available to reduce their expected losses. In particular, we study a setting in which employees can respond to incentives by exerting effort on the incentivized task, stealing from their employer and/or withholding voluntary effort on a different, nonincentivized 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 as a result of 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 (Harbring and Irlenbusch 2011, Neckermann et al. 2014, Bradler and Neckermann 2016).[2]

We study the impact of loss-framed incentives on these behaviors in a laboratory experiment in which 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, whereas participants in the *penalty* treatment start with an endowment and lose money

Downloaded from informs.org by [154.59.124.38] on 09 June 2023, at 05:30 . For personal use only, all rights reserved.

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 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 spillover* (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 up-front 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 nonincentivized behaviors by reducing the moral cost of selfish and unethical behavior—a phenomenon we refer to as the behavioral-spillover channel. There are various reasons why we might expect loss-framed incentives to cause behavioral spillover. First, they may provide "moral wiggle room" to justify unethical behavior (Rabin 1994, Dana et al. 2007, Mazar et al. 2008).[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 (Loewenstein 2000, Koszegi 2006, Gneezy and Imas 2014). 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 (Fehr and Schmidt 2006, Dur 2009).[5]

Whereas 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, whereas the behavioral-spillover 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-generating activities (i.e., task effort and theft) that decrease the relative incentives for survey completion. According to the behavioral-spillover 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. Whereas 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, 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

Downloaded from informs.org by [154.59.124.38] on 09 June 2023, at 05:30 . For personal use only, all rights reserved.

performance in various settings (e.g., Hannan et al. 2005, Brooks et al. 2012, Fryer et al. 2012, Hossain and List 2012, Armantier and Boly 2015, Levitt et al. 2016, DellaVigna and Pope 2017, De Quidt et al. 2017, Bulte et al. 2019). 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, Kern and Chugh 2009, Shalvi 2012, Grolleau et al. 2016, Pettit et al. 2016, Schindler and Pfattheicher 2017). Our study combines these two strands of the literature: we ask whether loss-framed incentives induce higher effort provision in a multitasking environment in which employees have access to both honest *and* dishonest ways of eliminating their losses.[7]

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. Whereas 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 nonincentivized 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, whereas 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 participant's 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, because of 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 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, envelopes, 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 cost functions: $c(s+\alpha z)$ captures the cost of effort to obtain score s and survey completion effort z (α is a

Downloaded from informs.org by [154.59.124.38] on 09 June 2023, at 05:30 . For personal use only, all rights reserved.

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 the participant completes the survey adequately, 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 nonpecuniary 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 $\kappa(t) = \bar{\kappa}_0 + \bar{\kappa}_{\mathbb{1}}(t)$ 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, whereas not stealing has no moral cost, stealing any amount (no matter how small) has a nonnegligible moral cost.[15]

Note that the utility function presented in Equation (1) represents participants' utility in the reward treatment. 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 zero if x is negative and x if x is nonnegative. This utility function generalizes the expression from Equation (1) in two dimensions. First, 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] Second, loss-framed incentives can activate the behavioral-spillover 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-spillover channel is inactive. If both channels are inactive, that is, when $\Lambda = 0$ and $\gamma = 1$, 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. Because behavior under loss-framed incentives is possibly driven by either the loss-aversion and/or the behavioral-spillover channels, we look at each channel's predictions for behavior. We 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 the optimization problem by choosing optimal task score s^*, amount of theft t^*, and survey completion effort z^*. Because 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 herein to help us to establish whether loss aversion or behavioral spillover is 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-spillover 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 incentive 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-spillover 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-spillover channel leads to a (weakly) higher share of participants for whom $t^* > 0$ in the penalty treatment as compared with the reward treatment. For the behavioral spillovers case, the result holds for t large enough, or under an additional assumption, as detailed in the proof in the appendix.*

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 with the total

Downloaded from informs.org by [154.59.124.38] on 09 June 2023, at 05:30 . For personal use only, all rights reserved.

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-spillover 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 see in Propositions 1 and 2, the loss-aversion and behavioral-spillover 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:

$$\max_{s,t,z} v(s+t) + \gamma rp(z) - c(s+\alpha z) - \gamma\kappa(t), \quad \text{s.t.} \quad s+t \geq R. \tag{2}$$

In other words, s^{**}, t^{**}, and z^{**} are the solutions to the maximization problem of an agent who always chooses combined income greater than 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-spillover 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), that is, $s^* \to s^{**}, t^* \to t^{**}$, and $z^* \to z^{**}$. As behavioral spillover becomes strong (i.e., $\gamma \to 0$), scores s^* and survey completion effort z^* converge to zero, 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 spillover is 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 is 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, for example, marathon runners attempting to finish below certain "round number" times (Allen et al. 2017) and taxpayers 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 only 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-spillover 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-spillover channel drives behavior.

3. Context and 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, and 320 individuals participated in the experiment. To ensure privacy and 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

Downloaded from informs.org by [154.59.124.38] on 09 June 2023, at 05:30 . For personal use only, all rights reserved.

in a different room throughout the experiment. The experiment consisted of computerized and pen-and-paper parts, 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 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.

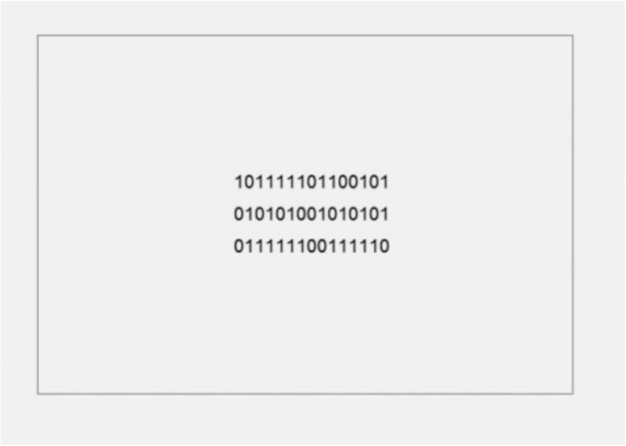
3.2. Design

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 three rows and 15 columns, and participants had 10 seconds per round to count how many zeros they contained.

After five 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 with 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-subjects 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 led 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.

Figure 1. Screenshot of the Computerized Matrix Task



```
101111101100101
010101001010101
011111100111110
```

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 nonincentivized 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 with 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

Downloaded from informs.org by [154.59.124.38] on 09 June 2023, at 05:30 . For personal use only, all rights reserved.

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 participant's cubicle and recorded whether any items were missing from the container after the experiment. Each box contained three 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€), and 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' 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 Endnote 16). This implies that we assume 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€ (which corresponds, therefore, to the initial status quo). This follows a well-established literature that considers the reference point is exogenously determined by the endowment or status quo (Tversky and Kahneman 1991, Ortoleva 2010, Masatlioglu and Ok 2014, Riella and Teper 2014).[24]

4. Data and 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 A.1). Table 1 provides the summary statistics of the demographic variables, the performance in the task, and participants' elicited opinions about the experiment.

Sixty percent of our participants are men, 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 one to seven. Even though participants do not find the experiment particularly fun (on average, they rate it 4.12 on a scale from one to seven), 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 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

Downloaded from informs.org by [154.59.124.38] on 09 June 2023, at 05:30 . For personal use only, all rights reserved.

Table 1. Summary Statistics

Variable (scale)	Mean	Standard deviation	Minimum	Maximum	*N*
Demographic variables					
Man (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

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 led 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 stole 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- versus loss-framed incentives by studying participants' combined income and satisfaction in the experiment.

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-spillover channel is ambiguous.

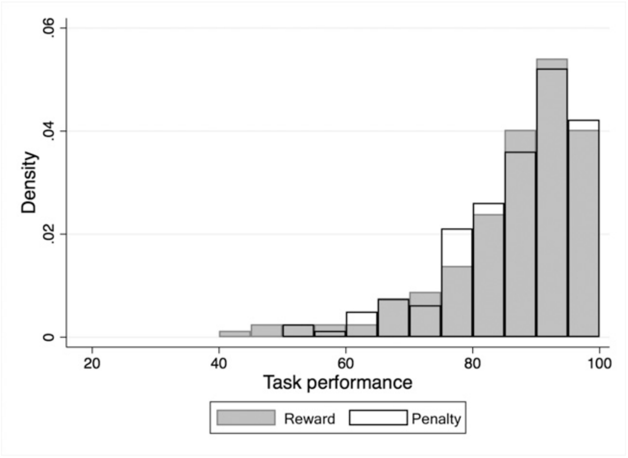
Figure 2 compares the distribution of total scores in the matrix task. In our study, there is no significant difference in performance between those who experience gain- versus loss-framed incentives: the group means are 85.7 and 86.2 in the reward and penalty treatments, respectively—a difference that corresponds to less than 1% of the mean in the reward treatment or approximately 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

Downloaded from informs.org by [154.59.124.38] on 09 June 2023, at 05:30 . For personal use only, all rights reserved.

Figure 2. Distribution of Task Scores by Treatment



margins when either the loss-aversion or behavioral-spillover channel is activated.

Our data show a clear impact of loss-framed incentives on theft. As can be seen in Figure 3(a), 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, whereas 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 3(b) presents the distribution of the value of stolen items and suggests that the penalty treatment disproportionately induces "small" theft.

5.3. Main Analysis: Survey Completion

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

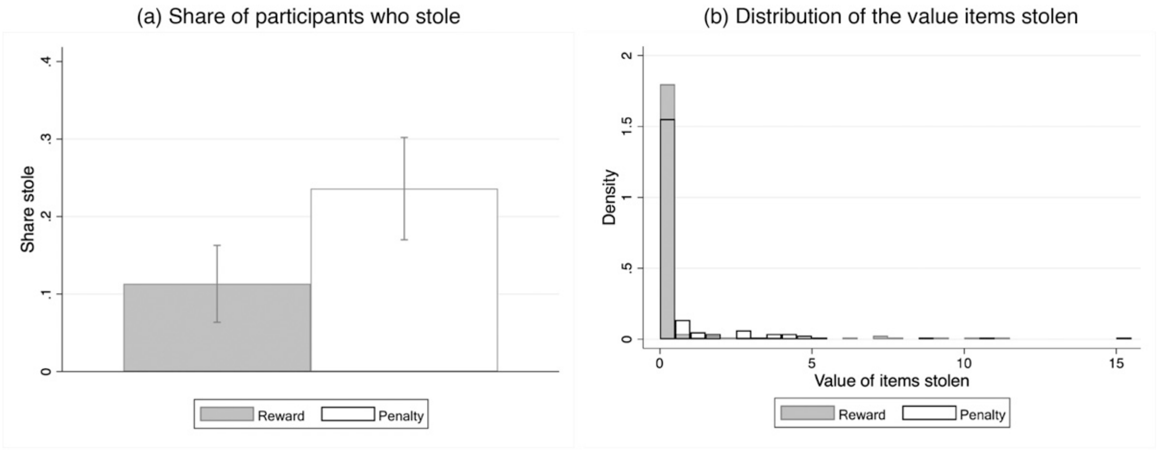
Figure 4 compares the share of participants who completed the voluntary survey between the two treatments. In the reward treatment 18.1% of participants complete the survey, whereas only 14.5% do so in the penalty treatment. This represents a reduction of 3.6 percentage points or 21.5%, which is not 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).

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 establishes that participants respond to loss-framed incentives. As we discuss in Section 2, we, however, need to go beyond comparative statics to

Figure 3. Theft by Treatment


(a) Share of participants who stole

(b) Distribution of the value items stolen

Downloaded from informs.org by [154.59.124.38] on 09 June 2023, at 05:30 . For personal use only, all rights reserved.

Figure 4. Survey Completion Rates by Treatment

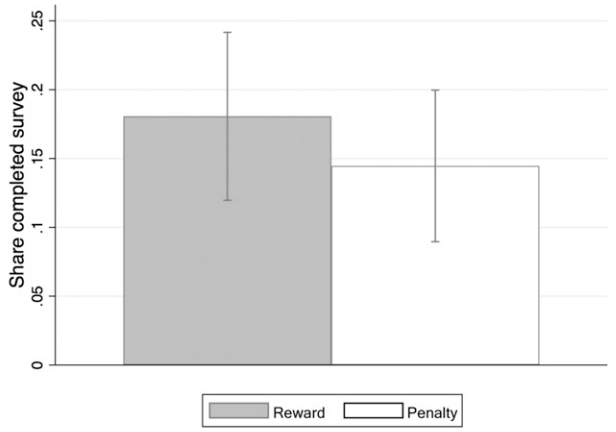
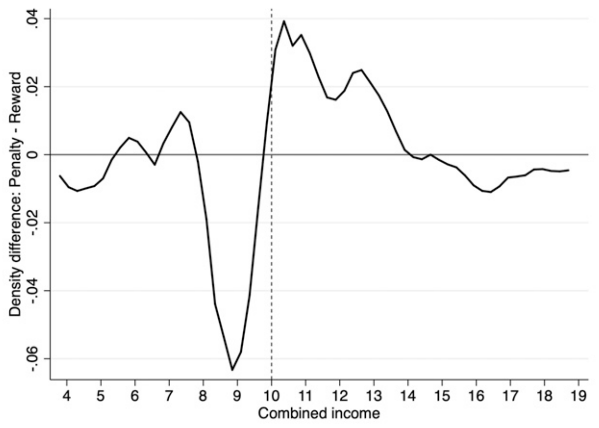


Figure 5. Difference in Kernel Density Estimates Between Penalty and Reward Treatments



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-spillover 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 estimate the kernel densities using 0.5 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€, that is, a shift of mass in the penalty treatment (as compared with the reward treatment) from the left of 10€ to the right of 10€.

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 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).

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

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)
	[0.006]	[0.011]	[0.020]	[0.458]	[0.290]
N	320	143	110	320	320

Notes. The table presents estimated marginal effects at the mean from probit models, in which 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. Randomization inference-based p-values in square brackets.

$^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Downloaded from informs.org by [154.59.124.38] on 09 June 2023, at 05:30 . For personal use only, all rights reserved.

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)) 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 drives 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 with 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 powered to detect an effect of similar size as observed in the studies mentioned.[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 behavioral-spillover channel. Recall that Proposition 1 predicts an ambiguous effect of loss-framed incentives on scores for the behavioral-spillover 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. Because 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 as a result of 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

Downloaded from informs.org by [154.59.124.38] on 09 June 2023, at 05:30 . For personal use only, all rights reserved.

present 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-spillover channel.[34] Second, answers in the obligatory questionnaire show that loss-framed incentives did not cause participants to express animosity toward the experimenter or question the fairness of their payment, suggesting that participants did not experience the negative emotions required to activate the behavioral-spillover channel. Given the strength of our framing intervention with the frequent reminders 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. Figures A.1 and A.2 show the distribution of guessed tasks scores and self-reported effort provision by treatment, whereas Figure A.3 compares the time spent on the calculation task per round by treatment. Mean *guessed* task score is slightly higher in the penalty than in the reward treatment (81.5 versus 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. 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 A.2 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 four instances when 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 findings 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 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. Although our real-effort task is certainly artificial, matrix tasks such as ours are used in many studies to mimic tedious jobs that require concentration (e.g., Abeler et al. 2011). Furthermore, even though 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.[35]

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 the way to becoming an employee, one might hope that our findings extrapolate to college-educated Western employees. A large body of research shows that loss aversion is an important

Downloaded from informs.org by [154.59.124.38] on 09 June 2023, at 05:30 . For personal use only, all rights reserved.

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, Kern and Chugh 2009, Shalvi et al. 2011, Grolleau et al. 2016, Pettit et al. 2016, Schindler and Pfattheicher 2017) and loss aversion and multitasking (Rubin et al. 2018, Pierce et al. 2020).

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 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% compared with 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 in which 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 artifact of the experimental nature of our study. Managers in real firms might be more likely to see negative behavioral spillover 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 needs to delve deeper into the study of various incentive 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.

Acknowledgments

The authors 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. The authors 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.

Appendix.

A.1. Proofs of the Theoretical Results

Proof of Proposition 1. We 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:

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

$$v'(s+t) - \gamma\kappa'(t) + \Lambda\, \mathbb{1}_{R>s+t} = 0, \tag{A.2}$$

$$\gamma r p'(z) - \alpha c'(s+\alpha z) = 0. \tag{A.3}$$

From Equation (A.3), we obtain

$$c'(s+\alpha z) - \frac{\gamma r}{\alpha} p'(z) = 0. \tag{A.4}$$

Table A.1. Balance Test

	Reward	Penalty	Difference
Man	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	

Notes. Comparison of means using *t*-tests with unequal variances. Age and year of study values are missing for 13 and 1 student(s), respectively.

$^*p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Downloaded from informs.org by [154.59.124.38] on 09 June 2023, at 05:30 . For personal use only, all rights reserved.

From Equations (A.1) and (A.2), we derive $c'(s+\alpha z) = \gamma\kappa'(t)$, and using that, in Equation (A.2), we obtain

$$\kappa'(t) - \frac{r}{\alpha}p'(z) = 0. \tag{A.5}$$

Equations (A.2), (A.4), and (A.5) jointly characterize the optimal s^*, t^*, and z^* for the case $t^* > 0$. For the problem with $t^* = 0$, the first order conditions are

$$v'(s) - c'(s+\alpha z) + \Lambda\mathbb{1}_{R>s} = 0, \tag{A.6}$$

$$\gamma rp'(z) - \alpha c'(s+\alpha z) = 0. \tag{A.7}$$

In what follows, for notational simplicity, we 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 (A.6) and (A.7). 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 (A.6) and (A.7), then we compute the Jacobian matrix with respect to s, z, and with respect to Λ,γ, respectively:

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

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

$$\begin{bmatrix} \frac{\partial s^*}{\partial \Lambda} & \frac{\partial s^*}{\partial \gamma} \\ \frac{\partial z^*}{\partial \Lambda} & \frac{\partial z^*}{\partial \gamma} \end{bmatrix} = -J_{s,z}^{-1} \times J_{\Lambda,\gamma} = -\frac{1}{det(J_{s,z})}\begin{bmatrix} \gamma rp'' - \alpha^2 c'' & \alpha c'' \\ \alpha c'' & v'' - c'' \end{bmatrix}$$
$$\times \begin{bmatrix} \mathbb{1}_{R>s} & 0 \\ 0 & rp' \end{bmatrix} =$$
$$= -\frac{1}{det(J_{s,z})}\begin{bmatrix} (\gamma rp'' - \alpha^2 c'') \cdot \mathbb{1}_{R>s} & \alpha rp'c'' \\ \alpha c'' \cdot \mathbb{1}_{R>s} & rp'(v'' - c'') \end{bmatrix}. \tag{A.8}$$

The determinant of $J_{s,t}$ is given by $det(J_{s,z}) = \gamma rp''v'' - c''(\gamma rp'' + \alpha^2 v'')$. Note that because of 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 (A.2), (A.4), and (A.5) and obtain the Jacobians with respect to s, t, z and Λ,γ, respectively:

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

We use once more the implicit function theorem to compute the comparative statics of s, t, z with respect to Λ and γ:

$$\begin{bmatrix} \frac{\partial s^*}{\partial \Lambda} & \frac{\partial s^*}{\partial \gamma} \\ \frac{\partial t^*}{\partial \Lambda} & \frac{\partial t^*}{\partial \gamma} \\ \frac{\partial z^*}{\partial \Lambda} & \frac{\partial z^*}{\partial \gamma} \end{bmatrix} = -J_{s,t,z}^{-1} \times J_{\Lambda,\gamma} =$$
$$= -\frac{1}{det(J_{s,t,z})}\begin{bmatrix} -\kappa''\left(\alpha c'' - \frac{\gamma rp''}{\alpha}\right) & \frac{rp''}{\alpha}(v'' - \gamma\kappa'') & (v'' - \gamma\kappa'')\left(\alpha c'' - \frac{\gamma rp''}{\alpha}\right) \\ \frac{rc''p''}{\alpha} & -\frac{rp''v''}{\alpha} & -v''\left(\alpha c'' - \frac{\gamma rp''}{\alpha}\right) \\ \kappa''c'' & -\kappa''v'' & -c''(v'' - \gamma\kappa'') \end{bmatrix}$$
$$\times \begin{bmatrix} \mathbb{1}_{R>s+t} & -\kappa' \\ 0 & -\frac{rp'}{\alpha} \\ 0 & 0 \end{bmatrix} =$$
$$= -\frac{1}{det(J_{s,t,z})}\begin{bmatrix} -\kappa''\left(\alpha c'' - \frac{\gamma rp''}{\alpha}\right)\mathbb{1}_{R>s+t} & \kappa'\kappa''\left(\alpha c'' - \frac{\gamma rp''}{\alpha}\right) - p'p''\frac{r^2}{\alpha^2}(v'' - \gamma\kappa'') \\ \frac{c''p''r}{\alpha}\mathbb{1}_{R>s+t} & -\frac{r\kappa'c''p''}{\alpha} + p'p''v''\frac{r^2}{\alpha^2} \\ \kappa''c''\ \mathbb{1}_{R>s+t} & -\kappa'\kappa''c'' + \frac{rp'\kappa''v''}{\alpha}. \end{bmatrix}.$$

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, because of 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 is negative if and only if $\kappa'\kappa''\left(\alpha c'' - \frac{\gamma rp''}{\alpha}\right) > p'p''\frac{r^2}{\alpha^2}(v'' - \gamma\kappa'')$. 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, that is, that it is not possible for a participant to go from $t^* > 0$ to $t^* = 0$. We start by proving the behavioral spillovers case, which becomes stronger as γ decreases. We consider a certain $\gamma_1 > 0$ and $\gamma_2 = \gamma_1(1-\epsilon)$. We denote by (s_j, t_j, z_j) the solutions to the first order conditions when $t > 0$ for γ_j and by $(\tilde{s}_j, 0, \tilde{z}_j)$ the solutions to the first order conditions when $t = 0$ for γ_j. We assume that the participant goes from $t^* > 0$ with γ_1 to $t^* = 0$ with γ_2 and reach a contradiction. From that assumption and the definition of (s_1, t_1, z_1) and $(\tilde{s}_2, 0, \tilde{z}_2)$, we have that $v(s_1 + t_1) + \gamma_1 rp(z_1) - c(s_1 + \alpha z_1) - \gamma_1\kappa(t_1) - \Lambda[R - s_1 - t_1]_+ \geq v(\tilde{s}_2) + \gamma_1 rp(\tilde{z}_2) - c(\tilde{s}_2 + \alpha\tilde{z}_2) - \Lambda[R - \tilde{s}_2]_+$ and $v(\tilde{s}_2) + (1-\epsilon)\gamma_1 rp(\tilde{z}_2) - c(\tilde{s}_2 + \alpha\tilde{z}_2) - \Lambda[R - \tilde{s}_2]_+ \geq v(s_1 + t_1) + (1-\epsilon)\gamma_1 rp(z_1) - c(s_1 + \alpha z_1) - (1-\epsilon)\gamma_1 \kappa(t_1) - \Lambda[R - s_1 - t_1]_+$. Adding both inequalities, we obtain

Downloaded from informs.org by [154.59.124.38] on 09 June 2023, at 05:30 . For personal use only, all rights reserved.

$r(p(z_1) - p(\tilde{z}_2)) \geq \kappa(t_1)$. However, $p(z)$ is a probability and, thus, bounded above by one, which means that the left-hand side of this inequality is bounded above by r, and thus, for $t_1 > \bar{t}$ with $\kappa(\bar{t}) \geq r$, it cannot hold (if we assume that $r < \bar{\kappa}_0$, then this inequality never holds), reaching the desired contradiction.

For the loss aversion case, suppose a participant chose $t^* > 0$ for a certain value $\Lambda_1 > 0$ and then $t^* = 0$ for a certain value $\Lambda_2 = \Lambda_1 + \epsilon$, and we show that there exists a contradiction.[36] We use the notation (s_j, t_j, z_j) for the solutions with $t > 0$ and Λ_j and $(\tilde{s}_j, 0, \tilde{z}_j)$ for the solutions with $t = 0$ and Λ_j. Notice that it must be the case that the original solution under Λ_1 must have had $s_1 + t_1 < R$ for, otherwise, the utility function would not have changed with the change in Λ. To show the contradiction, we consider two cases. First, we consider the case in which the participant strictly prefers the solution (s_1, t_1, z_1) to $(\tilde{s}_1, 0, \tilde{z}_1)$ for Λ_1. In that case, by the continuity of the relevant functions, for ϵ small enough, we must have that the participant prefers (s_1, t_1, z_1) to $(\tilde{s}_2, 0, \tilde{z}_2)$ for Λ_2, reaching a contradiction. The second case happens when the individual is indifferent between (s_1, t_1, z_1) and $(\tilde{s}_1, 0, \tilde{z}_1)$ for Λ_1. From the definition of s_1 and $\tilde{s}_2$, we have the following inequalities: $v(s_1 + t_1) + \gamma rp(z_1) - c(s_1 + \alpha z_1) - \gamma\kappa(t_1) - \Lambda_1(R - s_1 - t_1) \geq v(\tilde{s}_2) + \gamma rp(\tilde{z}_2) - c(\tilde{s}_2 + \alpha\tilde{z}_2) - \Lambda_1 (R - \tilde{s}_2)$, and $v(\tilde{s}_2) + \gamma rp \quad (\tilde{z}_2) - c(\tilde{s}_2 + \alpha\tilde{z}_2) - (\Lambda_1 + \epsilon)(R - \tilde{s}_2) \geq v(s_1 + t_1) + \gamma rp(z_1) - c(s_1 \ + \alpha z_1) - \gamma\kappa(t_1) - (\Lambda_1 + \epsilon)(R - s_1 - t_1)$. Adding both inequalities, we obtain $\tilde{s}_2 \geq s_1 + t_1$. However, from the fact that the individual is indifferent between (s_1, t_1, z_1) and $(\tilde{s}_1, 0, \tilde{z}_1)$ for Λ_1, it must be the case that $s_1 + t_1 > \tilde{s}_1$ and,[37] once again by continuity, we must have that, for ϵ small enough, $s_1 + t_1 > \tilde{s}_2$, reaching a contradiction. □

Proof of Proposition 3. We begin with the loss aversion case. Notice that the constrained problem is equivalent to maximizing the function $v(s+t) + \gamma rp(z) - c(s + \alpha t) - \gamma\kappa(t) - \mu(R - s - t)$, where μ is the Lagrange multiplier. Therefore, the first order conditions for this problem are identical to the first order conditions of the unconstrained problem except with μ instead of $\Lambda\mathbb{1}_{R>s+t}$. We have two cases. The first case happens when $s^{**} + t^{**} > R$ and $\mu = 0$, that is, when the optimal solution yields $s + t > R$ even for $\Lambda = 0$, in which case $(s^*, t^*, z^*) = (s^{**}, t^{**}, z^{**})$. In the second case, we have $s^{**} + t^{**} = R$, in which case the value of the Lagrange multiplier is given by $\bar{\mu} = c'(s^{**} + \alpha z^{**}) - v'(R)$ from conditions analogous to Equations (A.1) and (A.6). But then, for any $\Lambda \geq \bar{\mu}$, we must have $s^* + t^* = R$, and thus, $(s^*, t^*, z^*) = (s^{**}, t^{**}, z^{**})$. Hence, for Λ large enough, both solutions coincide as we want to show.

The proof for $\gamma \to 0$ is as follows. From the fact that $z \in [0,1]$ (a compact set) and $p'(z)$ is continuous, we have that $p'(z)$ is bounded and, therefore, that $\gamma \to 0$ implies $\gamma rp'(z) \to 0$. From this and Equation (A.3) (from the first order conditions), we have that $c'(s + \alpha z) \to 0$. Given that $c'(0) = 0$ and c is convex and, hence, c' is injective, we have that $s + \alpha z \to 0$, and because both s and z are nonnegative, that means $s^* \to 0$ and $z^* \to 0$. Now, from Equation (A.1), we have that, as $\gamma \to 0, v'(s + t) \to -\Lambda\mathbb{1}_{R>s+t}$. But note that, when $t > R$, $\mathbb{1}_{R>s+t} = 0$, and so we have that $v'(s+t) \to 0$, and therefore, $t^* \to \infty$ (because v is increasing and concave), which means that the participants' utility also converges to $\lim_{t\to\infty} v(t)$, and therefore, this is indeed the optimal solution, as we want to show. □

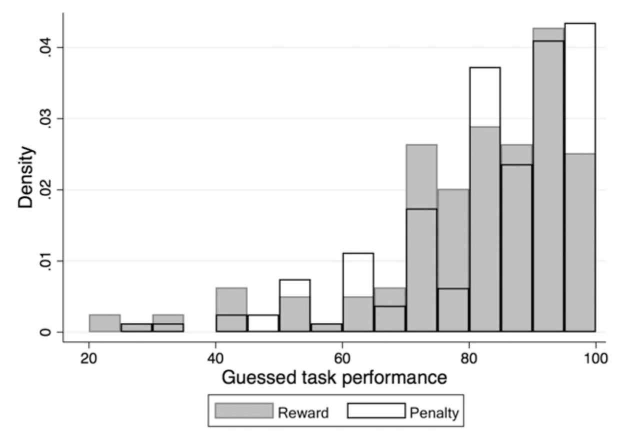
Proof of Proposition 4. Let s^*_R and t^*_R be the solutions for an agent i with cost functions (c_i, κ_i) in the reward treatment and s^*_P and t^*_P in the penalty treatment. We show first that there is more bunching in penalty at R when only the loss aversion channel is activated. Let $(c_i, \kappa_i) \in \mathcal{C}^+_{Reward}(\delta, R)$, 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 $\mathcal{C}^-_{Reward}(\delta, R) \not\subset \mathcal{C}^-_{Penalty}(\delta, R)$, we prove it by contradiction. Assume that $\mathcal{C}^-_{Reward}(\delta, R) \subset \mathcal{C}^-_{Penalty}(\delta, R)$, which means that, for all $(c_i, \kappa_i) \in \mathcal{C}$, whenever $s^*_R + t^*_R \in (R - \delta, R)$, then it must hold that $s^*_P + t^*_P \in (R - \delta, R)$. But, from Proposition 3, we know that, for Λ large enough, $s^*_P + t^*_P = s^{**} + t^{**} \geq R$, therefore reaching a contradiction.

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, we need to consider two cases. First, if $x > R$, then we can choose δ small enough such that $x - \delta > R$. But then, if $(c_i, \kappa_i) \in \mathcal{C}^-_{Reward}(\delta, x)$, it means that $s^*_R + t^*_R \in (x - \delta, x)$, and because $x - \delta > R$, we have $\mathbb{1}[R > s^*_R + t^*_R] = 0$, and irrespective of the value of Λ, we have $\Lambda\mathbb{1}[R > s^*_R + t^*_R] = 0$. This implies that 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 $\mathcal{C}^-_{Reward}(\delta, x) \not\subset \mathcal{C}^-_{Penalty}(\delta, x)$. Second, if $x < R$, from Proposition 3, we know that, as Λ becomes large enough, $s^*_P + t^*_P = s^{**} + t^{**} \geq R$; thus, for δ small enough, it is not true that $\mathcal{C}^+_{Reward}(\delta, x) \subset \mathcal{C}^+_{Penalty}(\delta, x)$, which concludes this part of the proof.

The argument to show that, under the behavioral spillovers channel, there is no point at which there is more bunching in penalty follows directly from Proposition 3 as, for any x and any $(c_i, \kappa_i) \in \mathcal{C}^+_{Reward}(\delta, x)$, we have that, as $\gamma \to 0$, $s^*_P + t^*_P \to +\infty$, and therefore, it is not true that $\mathcal{C}^+_{Reward}(\delta, x) \subset \mathcal{C}^+_{Penalty}(\delta, x)$. □

A.2. Figures and Tables for Additional Robustness Analysis

Figure A.1. Distribution of Self-Reported Guessed Performance by Treatment



Downloaded from informs.org by [154.59.124.38] on 09 June 2023, at 05:30 . For personal use only, all rights reserved.

Figure A.2. Distribution of Self-Reported Effort Provision by Treatment

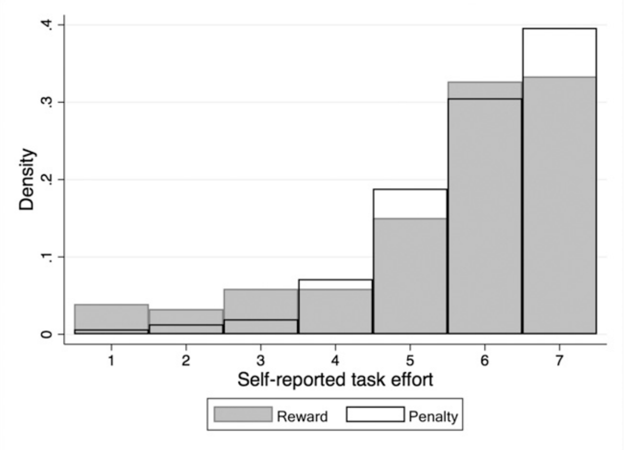
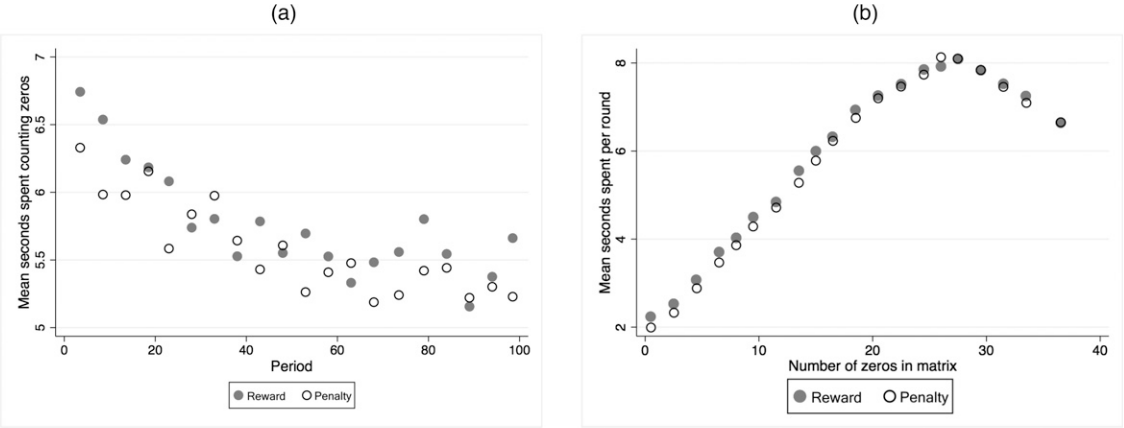


Figure A.3. Seconds Spent on Calculation Task Per Round



Notes. (a) Over time. (b) By task difficulty.

Table A.2. 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**

Endnotes

[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).

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

Downloaded from informs.org by [154.59.124.38] on 09 June 2023, at 05:30 . For personal use only, all rights reserved.

[4] Gneezy et al. (2011) provide numerous examples in support of the claim that the framing of a decision situation critically influences prosocial 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 employees become dissatisfied when their bonus payment falls below a natural reference standard for a fair bonus.

[6] The average task scores are 85.7 and 86.2 out of 100 in the reward and penalty treatments, respectively, a difference that corresponds to less than 1% of the mean in the reward treatment or approximately 5% of the pooled standard deviation. In the reward treatment, only 11.3% (18 out of the 159 participants) steal anything, whereas 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, whereas 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 versus 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, 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 nonincentivized task (matrix task versus voluntary survey), our design captures the idea that, whereas 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, whereas Rubin et al. (2018) find evidence for a quantity–quality trade-off in the laboratory. Studying financial incentives in the context of nonroutine, 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. Whereas there is a growing literature in behavioral economics on lying and cheating (e.g., Fischbacher and Föllmi-Heusi 2013, Kajackaite and Gneezy 2017, Gneezy et al. 2018), it is unclear to what extent these findings carry over to stealing (Belot and Schröder 2013, Hermann and Mußhoff 2019).

[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 noncash 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.

[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 because of, 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 for which the employee is not directly rewarded/punished.

[15] To assume a discontinuity at zero 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, theft of nonmonetary 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 participants 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 shows that people treat zero distinctly differently than positive amount. For instance, not paying subjects for a task can make them be more productive (because of 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 compare their score s with the reference point, then our model predicts that they steal *less* in loss- than in gain-framed incentives. This result arises because participants' utility function would then be $v(s+t)+\gamma rp(z)-c(s+\alpha z)-\gamma\kappa(t)-\Lambda[R-s]_+$, 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. Because our results indicate a substantial increase in the amount of theft, we abstain from discussing the case of subjects' only considering their task score.

[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 behavior when $s^*+t^*\geq R$.

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

[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 cannot 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 laboratory as time invested in order to benefit another participant.

Downloaded from informs.org by [154.59.124.38] on 09 June 2023, at 05:30 . For personal use only, all rights reserved.

[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. Online Figure A2 shows the boxes and the elements they contained.

[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, whereas 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 with perfectly anonymous situations in which 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 (Koszegi 2006, Barbos 2010, De Giorgi and Post 2011, Ok et al. 2015, Guney et al. 2018, Maltz 2020). Note that our setting can accommodate the existence of loss aversion in the reward treatment as long as the participant's reference point is close to the initial endowment of zero or at least below 5€. In both these cases, there would be no loss-aversion penalty.

[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 the 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.

[27] We also measured time spent on each matrix, the result of which is 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.

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

[30] The empirical literature on bunching (Chetty et al. 2011, Kleven 2016, Allen et al. 2017) 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.

[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, Hong et al. 2015, Grolleau et al. 2016, De Quidt et al. 2017, DellaVigna and Pope 2017). For a detailed review of the literature, please refer to De Quidt et al. (2017).

[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 note that the behavioral-spillover 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-spillover 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).

[35] It is unclear whether our treatment intervention with its frequent reminders was more or less strong compared with how loss-framed incentives are implemented in the field. We conjecture that there are fewer reminders in a typical field setting but that loss-framed incentives in the field are not less salient because of 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.

[36] We show this contradiction in a neighborhood of radius ϵ of Λ_1 for some $\epsilon > 0$, obtaining, thus, a local result. The global result is obtained from applying this reasoning to any $\Lambda_1 > 0$.

[37] The fact that $s_1 + t_1 > \tilde{s}_1$ for the second case can itself be proved by contradiction. If we had $\tilde{s}_1 \geq s_1 + t_1$, then $v'(\tilde{s}_1) \leq v'(s_1 + t_1)$ as v is concave and, thus, v' decreasing, and from Equations (A.1) and (A.6), we would have $c'(\tilde{s}_1 + \alpha\tilde{z}_1) \leq c'(s_1 + \alpha z_1)$. From Equations (A.3) and (A.7), this implies that $p'(\tilde{z}_1) \leq p'(z_1)$, and because p is concave and, thus, p' is decreasing, that means that $\tilde{z}_1 \geq z_1$. Thus, because $t_1 > 0$, this implies that $\tilde{s}_1 > s_1$ and $\tilde{z}_1 \geq z_1$, but this is a contradiction with the fact that $c'(\tilde{s}_1 + \alpha\tilde{z}_1) \leq c'(s_1 + \alpha z_1)$ as c is convex and, thus, c' increasing.

References

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

Allen EJ, Dechow PM, Pope DG, Wu G (2017) Reference-dependent preferences: Evidence from marathon runners. *Management Sci.* 63(6):1657–1672.

Armantier O, Boly A (2015) Framing of incentives and effort provision. *Internat. Econom. Rev.* 56(3):917–938.

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

Barberis NC (2013) Thirty years of prospect theory in economics: A review and assessment. *J. Econom. Perspect.* 27(1):173–196.

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

Belot M, Schröder M (2013) Sloppy work, lies and theft: A novel experimental design to study counterproductive behaviour. *J. Econom. Behav. Organ.* 93(3):233–238.

Belot M, Schröder M (2016) The spillover effects of monitoring: A field experiment. *Management Sci.* 62(1):37–45.

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

Bradler C, Neckermann S (2016) The magic of the personal touch: Field experimental evidence on money and appreciation as gifts. *Scandinavian Journal of Economics* 121(3):1189–1221.

Breza E, Kaur S, Shamdasani Y (2018) The morale effects of pay inequality. *Quart. J. Econom.* 133(2):611–663.

Brooks RRW, Stremitzer A, Tontrup S (2012) Framing contracts: Why loss framing increases effort. *J. Institutional Theoretical Econom.* 168(1):62–82.

Brownback A, Sadoff S (2019) Improving college instruction through incentives. *J. Political Econom.* 128(8):2925–2972.

Bulte E, List JA, Van Soest D (2019) Toward an understanding of the welfare effects of nudges: Evidence from a field experiment

Downloaded from informs.org by [154.59.124.38] on 09 June 2023, at 05:30 . For personal use only, all rights reserved.

in Uganda. NBER Working Paper 26286, National Bureau of Economic Research, Cambridge, MA.

Buser T, Dreber A (2016) The flipside of comparative payment schemes. *Management Sci.* 62(9):2626–2638.

Cameron JS, Miller DT (2009) Ethical standards in gain vs. loss frames. De Cremer D, ed. *Psychological Perspectives on Ethical Behavior* (Information Age Publishing, Charlotte, NC), 91–106.

Chetty R, Friedman JN, Olsen T, Pistaferri L (2011) Adjustment costs, firm responses, and micro vs. macro labor supply elasticities: Evidence from Danish tax records. *Quart. J. Econom.* 126(2):749–804.

Dana J, Weber RA, Kuang JX (2007) Exploiting moral wiggle room: Experiments demonstrating an illusory preference for fairness. *Econom. Theory* 33:67–80.

Danilov A, Vogelsang T (2016) Time for helping. *J. Econom. Sci. Assoc.* 2(1):36–47.

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

DellaVigna S, Pope D (2017) What motivates effort? Evidence and expert forecasts. *Rev. Econom. Stud.* 85(2):1029–1069.

De Quidt J (2017) Your loss is my gain: A recruitment experiment with framed incentives. *J. Eur. Econom. Assoc.* 51(5):351–365.

De Quidt J, Fallucchi F, Kölle F, Nosenzo D, Quercia S (2017) Bonus vs. penalty: How robust are the effects of contract framing? *J. Econom. Sci. Assoc.* 3(3):1–9.

Dur R (2009) Gift exchange in the workplace: Money or attention? *J. Eur. Econom. Assoc.* 7(2–3):550–560.

Englmaier F, Grimm S, Schindler D, Schudy S (2018) The Effect of incentives in non-routine analytical teams tasks—Evidence from a field experiment. CESifo Working Paper No. 6903, CESifo, Munich.

Fehr E, Schmidt KM (2006) The economics of fairness, reciprocity and altruism—experimental evidence and new theories. Kolm SC, Ythier JM, eds. *Handbook of the Economics of Giving, Altruism and Reciprocity* (Elsevier), 1:615–691.

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

Fischbacher U, Föllmi-Heusi F (2013) Lies in disguise—An experimental study on cheating. *J. Eur. Econom. Assoc.* 11(3):525–547.

Fryer R, Levitt S, List J, Sadoff S (2012) Enhancing the efficacy of teacher incentives through loss aversion: A field experiment. NBER Working Paper No. 18237, National Bureau of Economic Research, Cambridge, MA.

Gall T, Hu X, Vlassopolous M (2016) Dynamic incentive effects of team formation: Experimental evidence. IZA Discussion Papers 10393, Institute of Labor Economics (IZA).

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

Gneezy U, Imas A (2014) Materazzi effect and the strategic use of anger in competitive interactions. *Proc. Natl. Acad. Sci. USA* 111(4):1334–1337.

Gneezy U, Rustichini A (2000) Pay enough or don't pay at all. *Quart. J. Econom.* 115(3):791–810.

Gneezy U, Kajackaite A, Sobel J (2018) Lying aversion and the size of the lie. *Amer. Econom. Rev.* 108(2):419–453.

Gneezy U, Meier S, Rey-Biel P (2011) When and why incentives (don't) work to modify behavior. *J. Econom. Perspect.* 25(4):191–210.

Goette L, Huffman D, Meier S, Sutter M (2012) Competition between organizational groups: Its impact on altruistic and antisocial motivations. *Management Sci.* 58(5):948–960.

Gravert C (2013) How luck and performance affect stealing. *J. Econom. Behav. Organ.* 93(C):301–304.

Grolleau G, Kocher MG, Sutan A (2016) Cheating and loss aversion: Do people cheat more to avoid a loss? *Management Sci.* 62(12): 3428–3438.

Guney B, Richter M, Tsur M (2018) Aspiration-based choice. *J. Econom. Theory* 176(C):935–956.

Hannan RL, Hoffman VB, Moser DV (2005) Bonus vs. penalty: Does contract frame affect employee effort? Rapoport A, Zwic R, eds. *Experimental Business Research*, vol. II (Springer-Verlag, Berlin, Heidelberg), 151–169.

Harbring C, Irlenbusch B (2011) Sabotage in tournaments: Evidence from a laboratory experiment. *Management Sci.* 57(4):611–627.

Hermann D, 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. *J. Econom. Behav. Organ.* 163:135–139.

Holmstrom B, Milgrom P (1991) Multitask principal-agent analyses: Incentive contracts, asset ownership, and job design. *J. Law, Economics, Organization* 7:24–52.

Hong F, Hossain T, List JA (2015) Framing manipulations in contests: A natural field experiment. *J. Econom. Behav. Organ.* 118: 372–382.

Hossain T, List JA (2012) The behavioralist visits the factory: Increasing productivity using simple framing manipulations. *Management Sci.* 58(12):2151–2167.

Hsee CK, Yu F, Zhang J, Zhang Y (2003) Medium maximization. *J. Consumer Res.* 30(1):1–14.

Imas A (2016) The realization effect: Risk-taking after realized vs. paper losses. *Amer. Econom. Rev.* 106(8):2086–2109.

Imas A, Sadoff S, Samek A (2017) Do people anticipate loss aversion? *Management Sci.* 63(5):1271–1284.

Kahneman D, Tversky A (1979) Prospect theory: An analysis of decision under risk. *Econometrica* 47(2):263–292.

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

Kern MC, Chugh D (2009) Bounded ethicality: The perils of loss framing. *Psych. Sci.* 20(3):378–384.

Kleven HJ (2016) Bunching. *Annual Rev. Econom.* 8:435–464.

Koszegi B (2006) Emotional agency. *Quart. J. Econom.* 121(1):121–155.

Levitt SD, List JA, Neckermann S, Sadoff S (2016) The behavioralist goes to school: Leveraging behavioral economics to improve educational performance. *Amer. Econom. J. Econom. Policy* 8(4): 183–219.

Loewenstein G (2000) Emotions in economic theory and economic behavior. *Amer. Econom. Rev.* 90(2):426–432.

Maltz A (2020) Exogenous endowment-endogenous reference point. *Econom. J. (London).* 130(625):160–182.

Masatlioglu Y, Ok EA (2014) A canonical model of choice with initial endowments. *Rev. Econom. Stud.* 81(2):851–883.

Mazar N, Amir O, Ariely D (2008) The dishonesty of honest people: A theory of self-concept maintenance. *J. Marketing Res.* 45(6): 633–644.

Neckermann S, Cueni R, Frey BS (2014) Awards at work. *Labour Econom.* 31(C):205–217.

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

Ok EA, Ortoleva P, Riella G (2015) Revealed (p)reference theory. *Amer. Econom. Rev.* 105(1):299–321.

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

Pettit NC, Doyle SP, Lount RB, To C (2016) Cheating to get ahead or to avoid falling behind? The effect of potential negative vs. positive status change on unethical behavior. *Organ. Behav. Human Decision Processes* 137:172–183.

Pierce L, Rees-Jones A, Blank C (2020) The negative consequences of loss-framed performance incentives. NBER Working Paper 26619, National Bureau of Economic Research, Cambridge, MA.

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

Podsakoff PM, MacKenzie SB, Paine JB, Bachrach DG (2000) Organizational citizenship behaviors: A critical review of the theoretical and empirical literature and suggestions for future research. *J. Management* 26(3):513–563.

Rabin M (1994) Cognitive dissonance and social change. *J. Econom. Behav. Organ.* 23(2):177–194.

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

Rubin J, Samek A, Sheremeta RM (2018) Loss aversion and the quantity-quality tradeoff. *Experiment. Econom.* 21(2):292–315.

Ruggeri K, Alí S, Berge ML, Bertoldo G, Bjørndal LD, Cortijos-Bernabeu A, Davison C, et al. (2020) Replicating patterns of prospect theory for decision under risk. *Nature Human Behav.* 4:622–633.

Schindler S, Pfattheicher S (2017) The frame of the game: Loss-framing increases dishonest behavior. *J. Experiment. Soc. Psych.* 69:172–177.

Schurr A, Ritov I (2016) Winning a competition predicts dishonest behavior. *Proc. Natl. Acad. Sci. USA* 113(7):1754–1759.

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

Shalvi S, Handgraaf MJJ, De Dreu CK (2011) Ethical manoeuvring: Why people avoid both major and minor lies. *British J. Management* 22(s1):S16–S27.

Thaler RH, Johnson EJ (1990) Gambling with the house money and trying to break even: The effects of prior outcomes on risky choice. *Management Sci.* 36(6):643–660.

Tversky A, Kahneman D (1991) Loss aversion in riskless choice: A reference-dependent model. *Quart. J. Econom.* 106(4):1039–1061.

Young A (2019) Channeling Fisher: Randomization tests and the statistical insignificance of seemingly significant experimental results. *Quart. J. Econom.* 134(2):557–598.

Downloaded from informs.org by [154.59.124.38] on 09 June 2023, at 05:30 . For personal use only, all rights reserved.