The Impact of Monetary Penalties on Prosocial Motivation: Unveiling Crowding-Out and Crowding-In effects

Rafael Teixeira

November 1, 2023

Abstract

Monetary penalties are employed across various contexts and formats to deter undesirable behaviors. In some cases, the introduction of penalties leads to a deterioration of the situation, known as crowding-out effects, while in others, individuals display a propensity to follow rules and respect penalties, potentially resulting in crowding-in effects. This article contrasts these opposing theories by exploring how monetary penalties influence prosocial behavior and by exemplifying how minor contextual changes lead to different outcomes. We perform an experiment using a modified dictator game, subjecting participants to two subtly distinct penalty conditions: a "fine" - imposed after the dictator takes money - and a "fee" - paid before taking money. Our findings reveal that penalties have heterogeneous impacts on participants. While some individuals take more money when facing a penalty (crowding-out effect), others abstain from taking money, even when they take large amounts without the penalty (crowding-in effect). At the aggregate level, the "fine" shows no significant impact on the amount taken, suggesting the penalty's ineffectiveness. However, the "fee" leads to a significant reduction, illustrating how minor changes lead to different effects. Finally, our study demonstrates that monetary penalties trigger shifts in social norms. These shifts can partially explain the observed crowding-out and crowding-in effects.

Keywords: Crowding-out effect, crowding-in effect, fine, framing effects, social norm

JEL classification:
A13, D91, C91, K42

[1]I would like to extend my thanks to Wendelin Schnedler and Fabian Bopp from the University of Paderborn for their seminar and fruitful discussions, as well as to Alexis Belianin, Alejandro Hirmas, Shaul Shalvi, Ro'i Zultan, George Lowenstein, Silvia Sonderegger, Chris Starmer, Sander Onderstal, Jan-Willem Stoelhorst, Alexander Vostroknutov, Simon Gächter, Silvia Sonderegger and others for their valuable comments and suggestions.

1 Introduction

Monetary penalties are frequently used to influence human behavior by attaching financial consequences to undesirable actions. Surprisingly, in some instances, penalties might lead to crowding-out effects, causing incentives to backfire and actually increasing the targeted behavior. An illustrative case study in a daycare center (Gneezy and Rustichini (2000a)) demonstrated this unexpected result when a fine for late pickups led to even more tardiness. Conversely, recent literature (e.g. Kimbrough and Vostroknutov (2016, 2018)) indicates that people often exhibit rule-following tendencies, in the sense that they are willing to forgo personal gains to comply with rules. In these cases, rules may lead to crowding-in effects, encouraging more prosocial behavior. These theories offer contrasting conclusions and show that, even for a tool that is so widely used, it is difficult to predict the actual behavioral impact of monetary penalties on behavior.

Penalties are applied in various ways and settings. For instance, environmental legislation frequently incorporates both fees and fines to deter undesirable behavior that harms the environment. The EU employs emission permits as fees for companies, granting them the right to emit a certain amount of greenhouse gases. Conversely, firms that violate environmental regulations or laws often face fines as a form of punishment. Both situations reflect an additional cost associated with specific behaviors, and traditional economic theory does not differentiate between the format and timing of when a penalty is imposed, except for considerations related to risk and timing preferences. However, Sunstein (2003); Eriksson, Strimling, Andersson, and Lindholm (2017); Falk and Szech (2013); Bowles (2016) observe that the framing of incentives and rules can affect behavior by triggering different moral responses. Yet, it has never been directly tested how different penalty formats might lead to different consequences, and it remains unclear which settings are more effective.

This paper analyzes the influence of monetary penalties on prosocial motivation using an experiment. Our broader aim is to identify specific behavioral consequences resulting from the implementation of penalties while also examining whether small changes in the context can trigger different behavioral responses.

Our first objective is to comprehensively understand the interplay between crowding-out effects (a reduction in prosocial behavior due to penalties) and crowding-in effects (an increase in prosocial behavior as a response to penalties) in a controlled within-subject setting. Our experiment allows us to directly pinpoint changes in social concerns across settings and gain a clearer understanding of potential crowding-out (or crowding-in) effects. This study contributes to the existing literature that has tested similar phenomena in a penalty setting (e.g., Frey and Oberholzer-Gee (1997); Kurz, Thomas, and Fonseca (2014); Kornhauser, Lu, and Tontrup (2020); Kimbrough and Vostroknutov (2016)) by observing individual changes instead of aggregate responses and attempts to reconcile the coexistence of crowding-in and crowding-out, which are contradictory theories and should not happen simultaneously.

Secondly, we aim to test the differences between the effects of two types of monetary penalties: one designed to mimic a fine and another designed to mimic a fee. We have created stylized versions of these two monetary penalties, narrowing down the difference to simple framing effects (with the same set of possible behaviors and outcomes across conditions). Our goal is to determine whether even small differences in market designs can lead to varying impacts on behaviors (e.g. Falk and Szech (2013)) but within the context of monetary penalties. This research provides initial evidence that the market design of a penalty might result in different behavioral consequences. Policymakers and other stakeholders can explore

these findings to design more effective strategies for preventing undesirable behaviors.

Finally, our aim is to comprehend the underlying mechanism driving behavioral changes. Specifically, we design our experiment to rule out possible explanations associated with social preferences (e.g., Fehr and Schmidt (1999); Andreoni and Miller (2002); Charness and Rabin (2002)) and strategic interaction (Bénabou and Tirole (2003, 2006); Janssen and Mendys-Kamphorst (2004)) by having a situation where it would be theoretically unlikely to observe crowding-out effects due to these potential factors. Meanwhile, we believe that the implementation of a monetary penalty triggers distinct social norms (e.g., Lane, Nosenzo, and Sonderegger (2023a)), consequently leading to changes in behaviors.

We use the terminology established by Bicchieri (2016, 2005) and Xiao and Bicchieri (2010), which categorizes social norms into two components: 'empirical expectations' (beliefs about what others do) and 'normative expectations' (beliefs about what others think we ought to do). Social norms have been pointed out as a plausible mechanism behind the crowding-out (in) effects, as explained by Capraro and Perc (2021); Ellingsen and Mohlin (2022); Frey and Jegen (2001). Bénabou and Tirole (2006); Gneezy, Meier, and Rey-Biel (2011) also suggest that a perceived entitlement to act in a certain way can be used to rationalize and clarify such behaviors. We adapt the methodology proposed by Krupka and Weber (2013) to partially capture this sense of entitlement. By analyzing the role of social norms and entitlement, we provide direct evidence that crowding-out and crowding-in effects are associated with shifts in social norms due to the introduced new incentive.

To explore the behavioral consequences and the impact on prosocial motivation in response to monetary penalties, we analyze participants' choices in a modified dictator game. In this version of the dictator game, participants start with an initial endowment, and one participant (the dictator) can take money from another (the receiver). This act of 'taking money' symbolizes the undesirable behavior we aim to reduce through the penalty. Participants go through multiple rounds with different initial endowments in two different conditions, a control, and a treatment, in which a monetary penalty is introduced. In different groups, we implement two subtly different monetary penalties. In one treatment condition, participants face a penalty paid *after* the "unwanted" behavior, resembling a *fine*. In the other treatment condition, participants face a penalty paid *before* the "unwanted" behavior, mimicking a *fee*. We compare behavior across the treatments to understand how the format of the monetary penalty might lead to different outcomes. Lastly, we evaluate the social norms/entitlement in various scenarios across all conditions.

Our findings highlight the heterogeneous impacts of monetary penalties. In the "fine" condition, there is no notable impact on the aggregate amount taken, suggesting that this penalty may not be effective. In contrast, the "fee condition is effective and results in a significant reduction in the aggregate amount taken, with significant variations observed across the conditions. However, these aggregate results conceal two countervailing effects: a significant share of dictators who take points in the control condition do not do so when the penalty is implemented, thus decreasing the average amount taken. Conversely, dictators who continue to take money in the treatment conditions increase the amount taken, thereby raising the average. To further understand this result, we analyze changes at the extensive margin (if money was taken or not) and the intensive margin (conditional on money being taken, how much money was taken).

At the extensive margin, both the fine and fee conditions result in a reduction in the number of instances where money is taken. The fee condition leads to a significantly greater reduction than the fine. This additional decrease in the frequency of money being taken can partly explain the overall differences previously discussed between the conditions. More-

over, upon analyzing individual changes for those individuals who ceased taking money, we observed that many participants were taking larger amounts of money in the control conditions. In some cases, dictators would take all available money and, once the monetary penalties were introduced, they completely ceased taking any points. These substantial reductions provide evidence of a propensity to follow rules and indicate a crowding-in effect - a notable increase in the agents' prosocial concerns.

On the other hand, when analyzing the intensive margin and examining participants who continued taking money in the treatment condition, participants consistently took money more intensively in both the fine and fee conditions, with no significant differences across the conditions. This exemplifies a crowding-out effect, with participants becoming less socially concerned. This result is robust when controlling for individual differences and income effects.

When analyzing social norms as potential mechanisms behind behavioral change, we observe that the implementation of monetary penalties induces changes in these expectations. Participants, for example, believe that fewer individuals would be willing to take money with the implementation of penalties, but they also perceive taking large amounts of money as more socially appropriate when a penalty is in place. Intuitively, the logic is: "You should not do it, but if you do, you should make the most of it."

We also observe a significant positive relationship between empirical and normative expectations and decisions on both the extensive and intensive margins. This means that, for example, a participant is less likely to take money the fewer people she expects to take money. Moreover, we also find that social norms can partially account for the treatment effects in both the intensive and extensive margins, thereby partially explaining the crowding-out and crowding-in behaviors. However, social norms were unable to explain the differences between the fee and fine conditions, as the shifts induced by both are not significantly different across conditions.

The paper is structured as follows: Section 2 presents the experimental design, Section 3 includes the theoretical analysis and hypotheses. We begin with our experimental design and then move to the theory so we can more easily describe the theoretical predictions associated with the specific scenario that we construct. Section 4 contains the results, Section 5 discusses the implications of the findings, and Section 6 concludes.

2 Experimental Design

The experiment was conducted online using oTree (Chen, Schonger, and Wickens (2016)), and participants were recruited from Prolific. It lasted an average of 18 minutes, and participants earned an average of approximately £4.53, with 200 points equivalent to £1. All hypotheses, the experimental design, and regressions were pre-registered.[1]

We let participants interact in a dictator game in which the Dictator decides how much money to take from the Receiver. We modified the standard dictator game into a taking game to capture the impact of implementing a monetary penalty on an 'undesirable behavior.' The original dictator game incentivizes giving behavior, which is generally viewed positively. By reframing the game in terms of taking, we aimed to model a situation where such behavior is likely to be associated with 'stealing' or 'greediness.'

In the experiment, participants played a series of 20 dictator games. We used the strategic method, and all participants were asked to make decisions as if they assumed the role

[1] `https://osf.io/sqx38`

of the Dictator. They were informed that they would be randomly matched with another participant, and at the end of the experiment, they learned which role they had actually assumed: Participant 1 (the Dictator) or Participant 2 (the Receiver). One round was randomly selected, and participants received the amount chosen by the participant randomized as the Dictator. The payment was realized only at the end of the experiment, and the participants did not directly interact at any time.

In each round, the participants received an initial endowment, and they could decide how much money to take from the other participant. The initial endowments varied across rounds as shown in Table 1.

Cases	Dictator	Receiver
Behind - twins 1	100	800
	200	800
Behind - twins 2	170	730
	270	730
Behind - decoy 1	360	510
Ahead - twins 3	500	400
	600	400
Ahead - twins 4	550	350
	650	350
Ahead - decoy 2	630	310

Table 1: Cases - different initial endowments

The endowments are designed to create pairs: two pairs in which the Dictator starts with fewer points than the Receiver and two in which the Dictator starts with more points. In such pairs, the participants are always dividing a total of 900 or 1000 points, with the Dictator starting with an extra 100 points in the second case of each pair. This design is intended to create the possibility of controlling for income effects resulting from the introduction of the monetary penalty, which we will discuss after explaining the treatment conditions. To prevent participants from always choosing similar or equal numbers, two decoy cases with different numbers are included.

The experiment included a control condition and one of the two treatment conditions (fee and fine). Each session consisted of 10 rounds in the control condition and 10 rounds in the treatment conditions, with the order of the cases randomized. To check for a potential order effect associated with the treatment, different sessions started with different conditions, with some sessions starting with the treatment condition and others starting with the control condition.

Participants are presented with a box displaying the initial endowment, a slider to select the amount of money to take, and a confirmation button for their decision. In the treatment conditions, a 100-point monetary penalty is subtracted from the dictator if the dictator takes any amount greater than 0 points from the Receiver. The participants are informed about this change before the start of the 10 rounds with the penalty, and they are reminded of it on each decision screen in the treatment conditions. The specific text for each treatment condition can be found in Table 2.

Fee	In this round, there is a **price** of **100 points** to be paid **before 'taking'** any positive amount.
Fine	In this round, there is a **price** of **100 points** to be paid **after 'taking'** any positive amount.

Table 2: Text on each treatment

In the fine condition, the deduction of 100 points occurs after the participant has made their decision. Specifically, the participant selects the amount they would like to take, and if the chosen amount is greater than zero, 100 points are subtracted from the final outcome; otherwise, they retain their initial endowment.

In the fee condition, the deduction of 100 points occurs before the participant makes their decision. The participant is presented with the following question: "Would you like to pay 100 points to be able to take points from Individual 2?" If the participant chooses to pay the fee, 100 points are subtracted from their endowment, and the slider is activated, allowing them to decide on the allocation. If the participant chooses not to pay the fee, the slider remains blocked, and they are forced to take zero points.

The fee and fine conditions were designed to create stylized versions of their realistic counterparts, retaining key elements while making them as directly comparable as possible. We eliminated the risk and uncertainty typically associated with fines in such scenarios. If we included risk, we would have to adjust the penalty values. However, this would pose challenges in directly comparing fines with fees and in controlling for income effects, as it would lead to the creation of new endowments, and the values would not be consistent across the conditions.

On the other hand, the fee condition necessitates payment before the actual action, introducing a two-stage decision-making process that is absent in the control or fine conditions. In real-life scenarios, decisions in the absence of penalties (control condition) or with fines typically lack this aspect. The introduction of this element would directly impact the fine condition or complicate the comparison between fines and controls. Therefore, we retain this crucial element without negatively affecting the other conditions, as any fee in real life inherently encompasses both timing and commitment aspects.

We also made an effort to maintain consistent wording across conditions. For instance, we intentionally avoided using specific terms like "fee" and "fine" to minimize any potential moral burden associated with those words that could prime individuals and confound the analysis, making it challenging to disentangle the driving factors. This approach allows us to better assess behavioral changes and their underlying mechanisms.[2]

Notice that the values are the same for both fees and fines, creating merely a framing effect among the conditions. This framing effect becomes even more subtle when you consider that all payments are processed at the end of the experiment. Therefore, the only thing changing is the perceived timing of the payment.

When the penalty is introduced, and the participant still takes money, there is an efficient loss associated with the value of the penalty, subtracting 100 points from the total money available. Hence, some behavioral changes would be expected due to this income effect. Our twin cases are designed to control for this income effect.

Consider twin case 1, for example: In the control condition, the situation with 100/800 points can be directly compared to the 200/800 points in the treatment condition. When the

[2]Future research will further explore the impact of wording and the role of risk in such setting.

agent pays the fee or fine, subtracting 100 points, it reverts to the 100/800 scenario. Hence, at this juncture, all sets of possible outcomes are identical, and the decision should be the same. This means that to control for the income effect, we only utilize approximately half of the observations. Specifically, we compare the twin case with fewer points originating from the control condition to the one with more points originating from the treatment condition.

After all rounds of the dictator game, we elicit two potential mechanisms: social norms (including empirical and normative expectations) and entitlement. To do so, we asked participants to report their perceptions of entitlement, empirical expectations, and normative expectations for five cases (twins 2 (behind), twins 5 (ahead), and decoy 1). For each possible mechanism, one case was randomly selected for payment. Participants could earn an additional 100 points if their answers matched the group average. To maintain consistency and avoid confusion across the measures, we employed a linear rule to determine points earned based on the distance from the correct answer for all measures.

We assess how social norms and entitlement affect two types of behavior: whether the participants took any amount of money (the extensive margin) and how much money they took (the intensive margin).

To elicit empirical expectations, participants are asked to estimate the proportion of a hypothetical group of 100 participants who would take money in the dictator game. Subsequently, they are asked to provide an estimate of the average amount of points taken by those participants.

To elicit normative expectations, we used a questionnaire similar to the one developed by Krupka and Weber (2013) that evaluates appropriateness as judged by others through a coordination game. Participants rated different behaviors on a scale of 1 (very socially inappropriate) to 5 (very socially appropriate). The questionnaire aimed to capture the perceived normative expectations by asking participants to consider how others would evaluate what people ought to do in this situation. One question assessed the appropriateness of taking points (extensive margin), and the other question assessed the appropriateness of taking a significant amount of points (intensive margin), around 70% of the total (initial endowment + amount taken).

We use the same framework as Krupka and Weber (2013) and the coordination game to create a new measure for entitlement. While Krupka and Weber (2013)'s methodology is typically used to measure and incentivize appropriateness associated with a behavior, we adapt it to measure the social perception associated with perceived entitlement. To do that, we modify the question from "According to the other participants, how appropriate is it to take points in this situation?" to "According to the other participants, is Participant 1 entitled to take points in this situation?". We also changed the rating scale from 1 - Not entitled - to 5 - Completely entitled.

It is challenging to measure entitlement as it is a personal feeling that cannot be directly compared, and therefore, it cannot be directly incentivized. However, this personal feeling has a strong social component, as people need to perceive that they have permission to act in specific ways. Hence, by adapting Krupka and Weber (2013), we attempt to develop a format to partially capture this perception in an incentivized way and directly test theories that suggest changes in perceived entitlement can lead to crowding-out effects.

We also recorded the demographic information provided by Prolific, along with measures of positive reciprocity, negative reciprocity, trust, and altruism (Falk et al. (2018)), as well as a reactance scale (Hong and Faedda (1996)), which is a psychological measure associated with the level of conformity to rules and norms.

3 Theory and Hypotheses

This section is divided into four parts. The first part covers general literature on crowding-out and crowding-in effects. The second part links our experiment to models with social preferences (e.g., Fehr and Schmidt (1999); Andreoni and Miller (2002)) and incomplete information (e.g., Bénabou and Tirole (2006)), indicating that these models do not predict significant changes in behavioral responses between control and treatment conditions. The third part delves into social norms as potential mechanisms for behavioral changes, referencing models such as Krupka and Weber (2013) and Akerlof and Kranton (2000). The final part examines the distinctions between fees and fines.

3 .1 Crowding-Out and Crowding-in effects

Monetary penalties are often employed to influence behavior, with the aim of reducing undesirable actions. Rational choice theory describes that individuals and businesses evaluate the expected costs and benefits of their actions. As monetary penalties increase the cost associated with engaging in undesirable behavior, they have the potential to reduce such behavior (Becker (1968)).

However, incentives do not always align with rational choice theory. Titmuss et al. (1970) proposed that introducing monetary compensation for blood donation might reduce donations. This hypothesis was tested by Mellström and Johannesson (2008), revealing mixed results, including a decrease in blood donations among female participants when monetary rewards were offered. A similar study conducted by Frey and Oberholzer-Gee (1997) examined support for a nuclear waste storage facility and observed decreased support when monetary compensation was introduced. Gneezy and Rustichini (2000b) demonstrated that offering small monetary rewards led to reduced performance on various tasks, including logical exams. Similarly, Gneezy and Rustichini (2000a) reported that implementing a fine in a daycare for late-picking parents led to more late pickups.

These cases exemplify the crowding-out theory (e.g., Gneezy and Rustichini (2000a); Frey and Jegen (2001); Frey (2000); Frey and Oberholzer-Gee (1997)), which suggests that new extrinsic incentives may diminish prosocial concerns, leading individuals to act less prosocially. In our setting, similar to Gneezy and Rustichini (2000a), this theory implies that introducing a monetary penalty may increase the number of people taking points or the amount taken.

Conversely, rule-following behaviors, as described by Kimbrough and Vostroknutov (2016, 2018), suggest that people have rule-following tendencies even when they are against their monetary interests. For example, participants adhere to red traffic lights in simulations, even when it would be advantageous to disregard the rule.

In our context, a monetary penalty could be perceived as a new rule to follow, leading some participants to reduce the amount taken to conform to this new rule, potentially causing crowding-in effects and increasing prosocial motivation.

It is important to note that crowding-in and crowding-out theories describe opposing behaviors, with one indicating an increase in prosocial motivation and the other implying a decrease. Both theories cannot coexist within the same individual at the same time.

Either a penalty leads the individual to follow rules, comply, and engage in more prosocial behaviors, or the monetary penalty leads to more selfish behavior and a deterioration of the situation. The confrontation between these two possibilities is crucial for a better understanding of the impact of monetary penalties. In the following sections, we check the

models associated with such behavioral change in our experimental setting.

3 .2 Social Preferences and Crowding-Out(in) Effects

To analyze potential crowding-out and crowding-in effects in our experiment, we initially focus on two sets of models. Firstly, we analyze prosocial preferences models (e.g., Fehr and Schmidt (1999); Andreoni and Miller (2002); Charness and Rabin (2002)), which describe the utility function as relative gains across individuals and are generally used to explain behavior in dictator games. Then, we discuss models similar to those presented by Bénabou and Tirole (2006); Janssen and Mendys-Kamphorst (2004), which focus on the role of strategic interaction in explaining crowding-out behavior.

Consider an agent with a simplified version of inequality aversion (Fehr and Schmidt (1999)) participating in our experiment: the dictator with an initial endowment of x, and the receiver with an initial endowment of y. The dictator can take an amount of money, denoted as t, from the receiver, and β captures the level of inequality aversion. The agent's objective is to maximize:

$$U(x+t, y-t) = x + t - \beta|(x+t) - (y-t)|$$

Now, with the introduction of a penalty p, the agent has to maximize:

$$U(x+t, y-t) = \begin{cases} x - \beta|x-y| & \text{if } t = 0 \\ x + t - p - \beta|(x+t-p) - (y-t)| & \text{if } t > 0 \end{cases}$$

The agent, facing a penalty, has three options based on different β levels:

1. Indifference to inequality, taking everything.
2. Opting to maintain initial inequality due to efficiency loss ($-p$).
3. Minimizing inequality by redistributing the efficiency loss among participants, taking an extra $\frac{p}{2}$.

For illustration, consider an agent in a $(200, 800)$ scenario. In the control condition, someone with strong inequality aversion takes 300 points, resulting in a 500/500 split. Introducing the penalty, the agent loses 100 points, resulting in a $(100, 800)$ situation. Here, the agent takes 350 points, reaching a 450/450 taking 50 points more than before, a 'more intensive' choice that may be naively seen as a crowding-out effect.

To address this, our experiments use twin cases to control for income effect/efficiency loss. The $(200, 800)$ case in the treatment condition, after the penalty, becomes $(100, 800)$. $(100, 800)$ is the twin case associated with $(200, 800)$, sharing the same values given the agent pays the monetary penalty, and should yield the same decisions in the control condition: a 450/450 split with 350 points taken.

That is, in this experiment, each twin case can be compared as (x, y) and $(\hat{x}, y)$, where $\hat{x} = x + 100$. When the participant takes money in the treatment condition and pays the penalty ($p = -100$), we have $\hat{x} - p = x$, describing the same set of possible outcomes.

Considering twin cases, output-based models, like those used for prosocial preferences (e.g., Fehr and Schmidt (1999); Bolton and Ockenfels (2000); Andreoni and Miller (2002); Charness and Rabin (2002); Yang, Onderstal, and Schram (2016)), that focus on the final outputs in their utility function $U(x+t, y-t)$ would predict the same decisions when money

is taken in the treatment conditions.

Corollary 1: For twin cases (x, y) and $(\hat{x}, y)$, where $\hat{x} = x + p$, if $t^* > 0$, and $\operatorname{argmax} U(\hat{x} - p + t, y - t) = t^*$, then $\operatorname{argmax} U(x + t, y - t) = t^*$.

This implies that individuals who take money in the treatment condition should take the same amount in the control conditions when comparing their respective twin cases, making it not possible to observe any crowding-out or crowding-in effects.[3]

Similarly, several theories (e.g., Bénabou and Tirole (2006, 2003); Janssen and Mendys-Kamphorst (2004); Frey and Oberholzer-Gee (1997)) have attempted to explain crowding-out and crowding-in effects, highlighting two crucial aspects: incomplete information and strategic interactions.

In one possibility, a new incentive acts as a coordination tool in situations with multiple equilibria. For instance, people may perceive the penalty as new information indicating that many others are engaging in a specific behavior. In the absence of a known "correct equilibrium," the penalty leads individuals to coordinate their actions toward the same behavior (Janssen and Mendys-Kamphorst (2004)).

Another possibility is that the new incentive serves as a signaling tool for one's type. Individuals may engage in a certain action to portray themselves as a 'prosocial type' to others when others are uncertain about their type. However, when the new incentive increases the relative cost, the agent may stop such actions and act in accordance with their more 'selfish type' (Bénabou and Tirole (2006)).[4]

The dictator game exhibits two important characteristics that we leverage in our study: it minimizes the roles of strategic interaction and incomplete information. Participants have complete information about the outcomes, and there is no direct interaction with other participants. Moreover, the experiment is entirely anonymous, preventing participants or the experimenter from identifying who is performing specific actions.

Given these aspects, there is no opportunity for agents to use incentives to coordinate with others into different equilibria, and it is challenging to signal their types to others. Therefore, it would be unlikely to observe significant crowding-out and crowding-in effects between control and treatment conditions caused by those aspects.

Hence, we design our experiment in such a way that substantial behavioral changes are unlikely to be observed. If we do observe crowding-out or crowding-in behavior, it demonstrates that such changes in prosocial concerns can be observed even in the simplest settings.

Before moving forward, there is another aspect that requires further consideration: the agents who stop taking money due to the efficiency loss, $t^* = 0$.

Models such as those presented by Fehr and Schmidt (1999); Bolton and Ockenfels (2000); Andreoni and Miller (2002) offer similar insights about these agents. Individuals who take larger amounts for themselves are less likely to cease taking money. This is because when participants take larger amounts of money, they prioritize their self-interest over others and are less affected by the penalty[5]. If agents take larger amounts of money in the control

[3]Models for prosocial preferences have to rely on context-based utility to capture behavioral differences across the conditions. In the following section, we discuss how this context-based utility can be partially explained by potential changes in social norms.

[4]It is also possible to consider that the agent is sending signals to the experimenter or has self-image concerns. However, we believe that such aspects can be partially captured by social norms, as described later.

[5]In Appendix A, we provide an example of a quadratic function for inequality aversion to illustrate a

condition and stop taking money after the introduction of the penalty, these behavioral shifts can potentially be attributed to crowding-in effects.

We formulate our initial hypotheses based on classic economic theory models and those presented in this section. For the hypotheses, we also assume that we are comparing twin cases, which leads to clear and precise predictions. In the following section, we outline the alternative hypothesis and how we could observe crowding-out and crowding-in behavior in such a setting.

Hypothesis 1 - Aggregate Level: The introduction of the monetary penalty reduces the average amount taken by participants.

As Corollary 1 describes, in both the treatment and control conditions, agents take the same amount of points because they encounter identical potential outcomes. However, the introduction of a monetary penalty adds a fixed cost, which may discourage some agents from taking any points due to the associated efficiency loss.

As a result, the total amount taken is expected to decrease. This hypothesis can be extended to cover both the extensive margin (whether points were taken, or not) and the intensive margin (how many points were taken, if any).

Hypothesis 2 - Extensive Margin: The introduction of the monetary penalty reduces the proportion of cases in which participants take points.

Hypothesis 3: Participants who stop taking points when the penalty is introduced will tend to have taken low amounts in the control condition.

Hypothesis 2 suggests that some individuals will stop taking any points as the penalty introduces an efficiency loss. Hypothesis 3 examines the behavior of individuals taking larger amounts of money. As a benchmark, we compare the amounts taken with the size of the penalty, which is 100 points. Agents who take values close to 100 points in the control group are likely to be dissuaded from taking any amount of money after the penalty is introduced. However, agents taking larger amounts, for example, all the money available, should not stop taking money. Agents who were taking significantly more money than the monetary penalty and cease to take money when the penalty is implemented are considered potential crowding-in effects.

Hypothesis 4 - Intensive Margin: If a participant takes points in the treatment condition, there is no difference in the amount taken in the control and treatment conditions

Hypothesis 4 describes Corollary 1: given the twin cases, the set of potential outcomes is the same, and given that any money is taken, the amount should be the same.

3 .3 Social Norms and Perceived Entitlement

Psychological and economic theories (e.g., Frey and Jegen (2001); Capraro and Perc (2021)) suggest that incentives can influence intrinsic motivation and result in behavioral shifts. For

potential threshold for the amount of money that agents would take in the control condition and stop taking in the treatment condition.

instance, introducing a penalty can activate distinct moral and self-image concerns (e.g., Capraro and Perc (2021); Ellingsen and Mohlin (2022), Tonin and Vlassopoulos (2013)).

Similarly, rule-following behaviors and potential crowding-in effects, as emphasized by Kimbrough and Vostroknutov (2016), illustrate how individuals prioritize following norms and rules over personal gains. People conform to social norms and rules, viewing a monetary penalty as a new rule to follow, even when it does not benefit them individually.

Kimbrough and Vostroknutov (2016) describe that social preferences are primarily shaped by social norms, and the relation between social norms and prosocial behavior is supported by extensive literature (e.g., Bicchieri and Dimant (2019), Bicchieri (2005), Bicchieri (2016), Cialdini and Trost (1998)). Models such as Krupka and Weber (2013) and Akerlof and Kranton (2000), as well as experiments like Xiao and Bicchieri (2010), demonstrate that individuals tend to conform to social norms, which can be directly integrated into their preferences.

We believe that introducing new incentives might trigger different social norms, influencing behavior and potentially resulting in crowding-out or crowding-in effects. In our context, our goal is to capture conformity to these norms, suggesting a positive relationship between perceived norms and behavior. Changes in perceived norms should correspond to changes in behavior.

Additionally, entitlement may be a factor explaining crowding-out effects (e.g., Bénabou and Tirole (2006); Gneezy et al. (2011)). To explore this, we adapted a measure from Krupka and Weber (2013) based on a coordination game to gauge group opinions.

Our methodology is also inspired by attribution theory from social psychology (e.g., Peterson et al. (1982); Dykema, Bergbower, Doctora, and Peterson (1996)), examining how individuals perceive causes and motivations behind experiences. This method partially captures the social construction of motivation (entitlement) based on context and individuals.

We illustrate this concept with a utility function that includes social norms. The agent's utility, U, depends on their initial endowment x, the amount taken t, the penalty p, and social norms $N(t, t^k_{\text{emp}}, t^k_{\text{nor}}, t^k_{\text{ent}})$. These norms are integrated into the utility function, where t^k_{emp} represents empirical expectations, t^k_{nor} represents normative expectations, and t^k_{ent} represents perceived entitlement. The norms are conditional on a specific condition, k. We also introduce a parameter γ, representing the agent's propensity to conform to norms:

$$U(t) = \begin{cases} x - \gamma N(0, t^k_{\text{emp}}, t^k_{\text{nor}}, t^k_{\text{ent}}) & \text{if } t = 0 \\ x + t - p - \gamma N(t, t^k_{\text{emp}}, t^k_{\text{nor}}, t^k_{\text{ent}}) & \text{if } t > 0 \end{cases}$$

We expect a positive relation between the amount taken t and social norms. To illustrate this numerically, we can use a model similar to Akerlof and Kranton (2000), incorporating empirical expectations, $E^k[t]$, into the utility function. In this case, the amount taken is related to empirical expectations through a quadratic formula:

$$U(t) = \begin{cases} x - \gamma(E^k[t] - 0)^2 \text{if } t = 0 \\ x + t - \gamma(E^k[t] - t)^2 \qquad \text{if } t > 0 \end{cases}$$

In this scenario, the amount taken, t^*, can be expressed as $t^* = E^k[t] - \frac{1}{2\gamma}$. Thus, higher empirical expectations lead to a larger amount taken. We will examine behavioral changes in the extensive and intensive margins. We also expect that different conditions would lead to different social norms. Based on this conformity and these shifts in the social norms, we establish the following hypotheses:

Hypothesis 5 - Norm Shifts: The implementation of monetary penalties impacts social norms (empirical, normative, and entitlement).

Hypothesis 6: Higher empirical/normative/entitlement values for taking any/larger amounts of money are associated with a higher likelihood/larger amounts of taking any money.

If the introduction of the monetary penalty affects social norms/entitlement (Hypothesis 5), and the agent conforms to social norms (Hypotheses 6 & 7), we can derive the following corollaries:

Corollary 2 - Crowding-Out Effect: For twin cases (x, y) and $(\hat{x}, y)$, where $\hat{x} = x+p$, if $t_i^{\hat{x}-p} \geq t_i^x$ for $i \in (emp, nor, ent)$, then $\text{argmax}\, U(\hat{x}-p+t, N(.)) = \hat{t}^* \geq t^* = \text{argmax}\, U(x+t, N(.))$.

Corollary 3 - Crowding-In Effect: For twin cases (x, y) and $(\hat{x}, y)$, where $\hat{x} = x+p$, if $t_i^{\hat{x}-p} \leq t_i^x$ for $i \in (emp, nor, ent)$, then $\text{argmax}\, U(\hat{x}-p+t, N(.)) = \hat{t}^* \leq t^* = \text{argmax}\, U(x+t, N(.))$.

Similar arguments can also be applied to the likelihood of taking any money (the extensive margin).

In summary, people tend to conform to social norms in a positively monotonic manner. If monetary penalties affect social norms, behaviors will reflect these changes. If the penalty negatively affects the social norm, behavior will deteriorate, leading to crowding-out effects. If the penalty positively affects the social norm, behavior will improve, resulting in crowding-in effects.

3 .4 Fine vs. Fee

Monetary penalties and fines serve diverse purposes, from managing parking issues to ensuring environmental regulation compliance. Although they share an economic objective of increasing the cost of certain behaviors, their formats can lead to distinct psychological interpretations, as discussed by Sunstein (2003) and Bowles (2016).

To examine potential behavioral changes resulting from altering the penalty format, we introduce two stylized versions: a fee (paid before taking money) and a fine (paid after taking money).

Importantly, we exclude a risk component in the fine, making it fundamentally different from real-life scenarios. This adjustment is made to ensure identical outcomes for fees and fines, allowing for a clear and direct comparison.

The fee introduces a two-stage decision that is absent in the fine. Implementing a fee without this initial decision step is not feasible. In contrast, the typical implementation of fines and control conditions in a more realistic context lacks this first-stage decision. We consider this distinction fundamental across conditions. Subsequent research will aim to disentangle this factor from the payment itself. However, our current focus is to explore potential variations between fee and fine versions, capturing essential distinctions while controlling for factors that could hinder precise case or theory comparisons. We emphasize this difference as a key driver of variation across conditions, reflecting how these cases might be different in real-life settings.

It is important to emphasize that both fees and fines impose a 100-point cost with the same economic impact. In traditional economic theory, both would yield identical outcomes and should lead to the same behaviors, as described in section 3.2.

However, the different timing and perceived commitment could lead to distinct behavioral changes.

The fee's upfront payment might highlight the moral salience (e.g., Bordalo, Gennaioli, and Shleifer (2013)) of the decision (e.g., Eriksson et al. (2017)), potentially reducing the number of people choosing to take money, suggesting a greater crowding-in effect for fees.

Conversely, the concept of entitlement discussed by Gneezy and Rustichini (2000a) and Gneezy et al. (2011) could contribute to a crowding-out effect with fees. Participants might feel they have the right to act upon what they've already paid for.

Hence, different theories suggest that fees could result in larger crowding-out effects and stronger adherence to rules.

Our aim is to test whether fees and fines lead to different social norms and, consequently, different types of behavior, in line with the literature on framing effects. Ellingsen, Johannesson, Mollerstrom, and Munkhammar (2012) and Krupka and Weber (2013) show how framing can influence behavior by affecting empirical and normative expectations. Thus, similar to what we observe in these articles, we understand that differences in moral perceptions might be reflected in different social norms between fees and fines.

We will investigate behavioral differences between fees and fines across all previous hypotheses, which we collectively refer to as "Hypothesis B." In summary, Hypothesis B encompasses:

Hypothesis B: There are behavioral differences (in terms of the amount of money taken and the likelihood of taking money) and differences in social norms (empirical/normative/entitlement) between fees and fines.

There could be other potential causes for differences between fees and fines. For instance, various payment methods can elicit different responses, even when the amounts involved are the same, as demonstrated by Zellermayer (1996). By making the payment more salient in the first stage of the fee condition, people might be less likely to take money due to the increased emotional burden. Similarly, Read, Loewenstein, Rabin, Keren, and Laibson (2000) describes that people act differently when facing narrow or broad bracketing, i.e., when considering the problems separately or together. Such cognitive and emotional aspects can be associated with differences between fees and fines, without necessarily triggering other social norms.

4 Results

The study involved 201 participants, with 101 in the fee condition and 100 in the fine condition, resulting in 4020 decisions. In the first part of the analysis, we describe the behaviors for all the observations, but we primarily focus our main analysis on the twin cases, totaling 1608 observations, that account for income effects. Additionally, participants provided information on social norms and perceived entitlement for one twin case where the dictator is behind (twins 2) and one where the dictator is ahead (twins 4), resulting in 804 observations for each case.

We checked for order effects, as different sessions started with either control or treatment

conditions. We observed no significant difference across the order, as shown in Appendix 10, and thus, all the corresponding treatment sessions are grouped together for data analysis. Appendix 9 demonstrates that the groups are balanced between conditions, with similar age, gender, and ethnicity.

The findings are presented in two sections. Section 4.1 explores the effect of monetary penalties on taking behavior, analyzing overall changes and breaking it down into the extensive and intensive margins while examining the behavioral differences between fines and fees. In section 4.2, the study examines the role of social norms and entitlement in the amounts taken by participants and analyzes these changes as potential behavioral explanations.

4 .1 Changes in the prosocial behavior

We start by investigating the impact of the monetary penalties on aggregate behavior using the following regression equation:

$$Take_{i,r} = \beta_0 + \beta_1 Fine + \beta_2 Fee + \beta_3 ControlDiff + \epsilon_{i,r}$$

Our aim is to explain the amount taken (*Take*) by individual i in round r. β_0 captures the mean behavior of the control condition in the fine treatment. The variable *Fine* is a dummy for the fine treatment, and β_1 captures the fine treatment effects. *ControlDiff* is a dummy for all sessions with the fee condition, and β_3 captures any potential differences for the control conditions across the different treatment conditions[6]. *Fee* is a dummy for the fee treatment, and β_2 captures the fee treatment effects. After running the regressions, we perform a chi-square test comparing β_1 and β_2 to check if the fee and fine have different impacts. We use a random effects model to control for individual differences, and the residuals are clustered at the individual level.

Table 3 presents the results of the regression analyses for the aggregate impact of each treatment. Regression 1 displays the impact when considering all data, and Regression 2 focuses on the twin cases:

[6]This coefficient serves as a robustness check for balance across the sessions at the aggregate level; however, it also has another interpretation on the intensive margin, as will be discussed.

	(1) Take	(2) Take
Fine	5.426	-6.163
	(6.448)	(7.614)
Fee	-23.35***	-27.78***
	(8.300)	(10.19)
ControlDiff	-5.289	-5.123
	(20.53)	(21.55)
Constant	283.2***	317.4***
	(15.24)	(15.82)
N	4020	1608

Standard errors clustered at the individual level in parentheses

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

Table 3: Aggregate treatment effects on the amount taken - (1) all observations, (2) twin cases

The results between regression (1) – without controlling for income effects – and regression (2) – controlling for the income effect – are very similar. We consider regression (2) as our primary benchmark. Notably, a statistically significant decrease in the amount taken is observed in the fee condition (-27), providing support for Hypothesis 1. Conversely, the fine condition shows a non-significant decrease (-6). A comparison of the fee and fine treatments reveals a marginally significant difference ($\chi^2(1) = 2.89, p = 0.0894$).

Result 1: *The implementation of a* ***fee*** *led to a significant reduction in the amount taken, while the implementation of a* ***fine*** *did not result in a significant change.*

Result 1B: *There are marginally significant differences between the* ***fee*** *and the* ***fine****, with the* ***fee*** *leading to a significantly lower amount taken compared to the fine.*

To gain a deeper understanding of these behavioral shifts, we can closely examine the distribution of individual changes in behavior within both the control and treatment conditions. That is, we check the average difference between the twin cases within both the control and treatment for each individual given each condition, as illustrated in Figure 1.

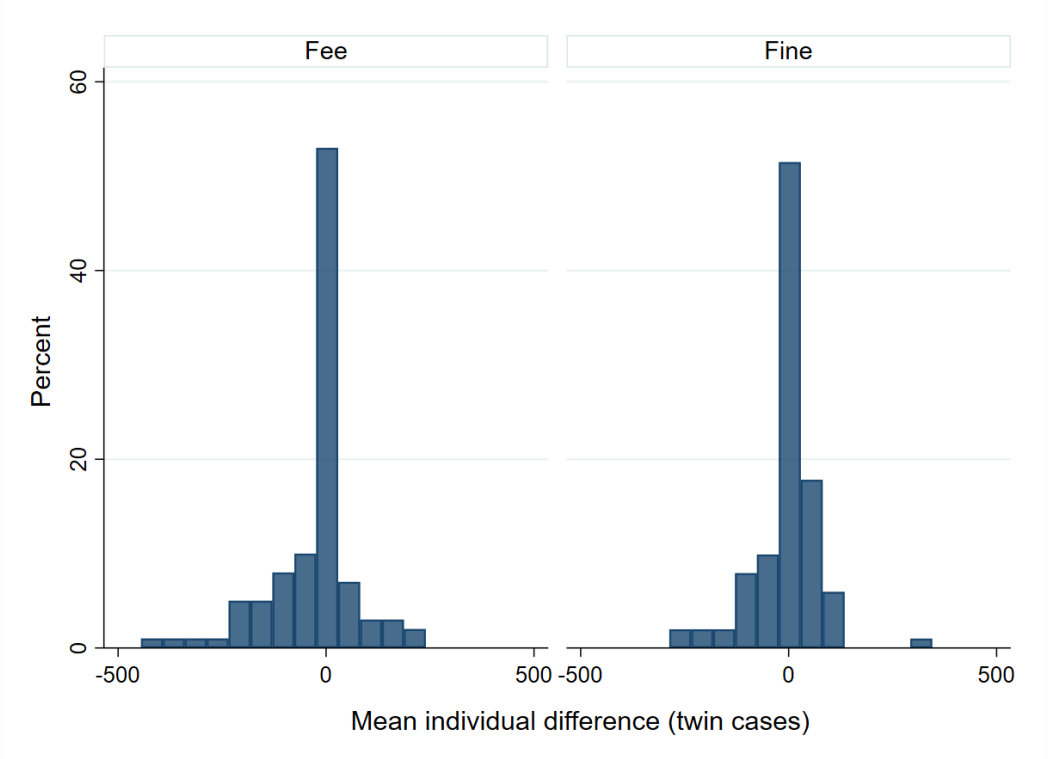

Figure 1: Distribution of mean individual changes for each condition

The behavior of the majority of participants remains relatively consistent, taking the same amount of money in both treatment and control conditions. However, we also observe significant deviations among some individuals. For instance, one participant reduced their total by -435 points, while others substantially increased their takings by up to 235 points.

To gain deeper insights into these individual behavioral changes, we examine two contrasting effects: some individuals completely stop taking points (extensive margin), while others demonstrate significant changes in the intensity of their takings (intensive margin). This sets the stage for our subsequent hypotheses. We begin by analyzing the extensive margin, as outlined in Hypothesis 2, which suggests that monetary penalties should result in a reduction in the number of participants taking money.

To analyze behavioral changes on the extensive margin, we perform a regression similar to the previous one. However, we modify the dependent variable to a binary outcome, "Participation," which equals one if money was taken and zero otherwise. Additionally, we employ a logit regression with random effects. Table 4 presents the results, with Regression (3) using the entire dataset, and Regression (4) focusing on the twin cases.

	(3) Participation	(4) Participation
Fine	-0.514***	-0.388**
	(0.139)	(0.156)
Fee	-1.269***	-0.962***
	(0.159)	(0.159)
ControlDiff	0.142	0.0343
	(0.294)	(0.258)
Constant	1.902***	1.712***
	(0.214)	(0.200)
N	4020	1608

Standard errors clustered at the individual level in parentheses

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

Table 4: Extensive margin: the impact of the fee and fine on the number of cases where money is taken - (3) all observations, (4) twin cases

The observations provide evidence that supports Hypothesis 2 for both regression (3) and (4). Using regression (4) as our main benchmark, there is a decrease in the percentage of cases where points are taken in both the fee and fine conditions (80.19% vs. 64.64% and 80.65% vs. 75.06%, respectively, for the twin cases).

We conducted a chi-square test to analyze the 10-percentage-point difference in impacts between the fee and fine treatments ($\chi^2(1) = 5.01, p = 0.0252$). The results indicate significant differences between the fee and fine treatments.

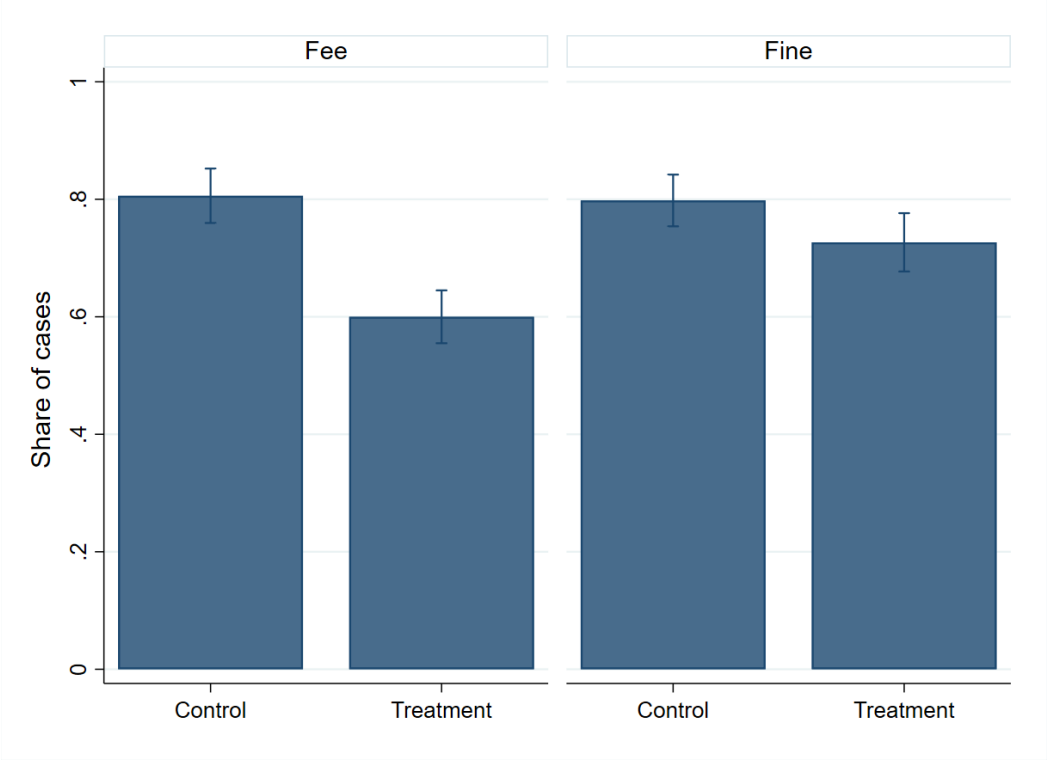

Figure 2: Predicted value of the share of cases in which money is taken and their 95% confidence interval in each condition (Twin Cases).

Considering that individuals are similar across the conditions, this increase in the number of cases in which money is taken can be associated with a crowding-in effect linked to the fee relative to the fine condition.

Result 2: *The implementation of both the* ***fee*** *and the* ***fine*** *leads to a significant reduction in the percentage of cases in which points are taken.*

Result 2B: *Significant differences exist between the* ***fee*** *and the* ***fine****, with the* ***fee*** *resulting in an even larger reduction in the percentage of cases compared to the* ***fine****.*

As described in the theory section, it is expected that some agents would stop taking money. We investigated the behavior of those agents during the control condition, i.e., how much they took in the control condition for the twin case in which they ceased taking money in the treatment condition. Figure 3 shows the distribution of the share[7] that the dictator earned in the control condition for situations in which no money was taken in the treatment condition, while Figure 4 displays the distribution of the amount taken for the same respective cases.

7 $\frac{\text{take+endowment}}{\text{total}}$

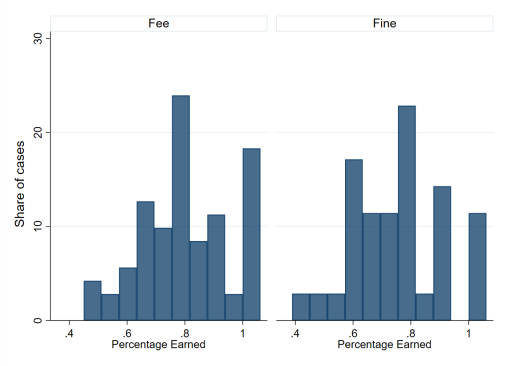

Figure 3: Distribution of the share earned by the dictator in the control condition for those who did not take money in the treatment condition.

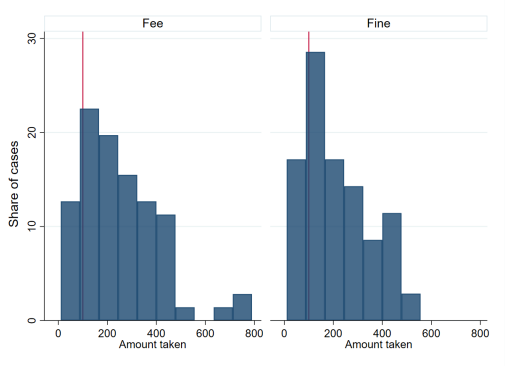

Figure 4: Distribution of the amount taken by the dictator among those who did not take money in the treatment condition.

On average, dictators obtain around 80% and 77% of the total available in the fee and fine conditions, respectively, for those specific cases. In some instances, the dictator acquires larger amounts or even the entire available amount but then ceases to take any money after the penalty's implementation. More precisely, in the fee condition, dictators obtain 100% of the money in 18.3% of cases, while in the fine condition, this occurs in 11.43% of cases, and these individuals decide to stop taking any money after the penalty is imposed.

Additionally, we compare the amount taken to directly assess it against the 100-point monetary penalty. Among those cases, the fee results in an average reduction of 248 points, whereas the fine condition shows a reduction of 200 points, with no significant differences between the treatment conditions ($\chi^2(1) = 0.88, p = 0.3482$). As a benchmark criterion, we compare the amount taken with the 100-point cost of the monetary penalty, and the average amount taken is significantly different ($\chi^2(1) = 42.50, p = 0.0000$). To illustrate further, in approximately 50% of the cases, participants take more than 200 points, and in around 30% of the cases, they take more than 300 points but then cease taking money in the treatment conditions. These substantial reductions in the amount taken provide evidence of a crowding-in effect.

Result 3: *Among dictators who did not take any money in the treatment condition, implementing either the* ***fee*** *or the* ***fine*** *led to reductions that were significantly larger than the 100 points associated with the monetary penalty.*

Result 3B: *There was no significant difference between the* ***fee*** *and* ***fine*** *conditions among dictators who did not take any money in the treatment condition.*

We now proceed with the intensive margin analysis. Before beginning, we must clarify the sample used in each subsequent regression. In general, the intensive margin focuses on participants who took any money, as represented by regression (5). However, it is expected that the participants who took money in the treatment and control conditions may differ, potentially introducing an endogenous effect due to different individuals in each condition.

To control for this aspect, we specifically analyze the amount taken by participants who

took money in the treatment condition and compare it with their corresponding control case, ensuring the same participants and cases for both conditions. Regression (6) presents the results for the entire dataset, while regression (7) focuses on the twin cases, controlling for individual and income effects.

Notice that the coefficient, *ControlDiff*, is intended to capture whether the participants willing to take money after the fee or fine conditions significantly differ. If this is the case, *ControlDiff* will account for these differences. Table 5 offers additional details:

	(5) Take	(6) Take	(7) Take
Fine	38.66***	35.67***	15.45**
	(6.592)	(6.657)	(7.539)
Fee	78.63***	37.22***	25.31***
	(8.817)	(6.795)	(8.754)
ControlDiff	1.505	-38.42**	-26.93
	(16.24)	(17.72)	(19.45)
Constant	338.8***	384.3***	417.8***
	(12.19)	(13.91)	(15.16)
N	2946	2668	1118

Standard errors clustered at the individual level in parentheses

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

Table 5: Intensive margin: The impact of the fee and fine on share earned by the dictator conditional on taking money in the treatments - (5) all observation, (6) twin cases

The results contradict hypothesis 3, suggesting increases in the amount taken, and we observe crowding-out effects for all regressions.

After controlling for income effects, regression (7), both the fee and fine conditions lead to a significant increase in the amount taken (15.45 and 25.31, respectively). We conducted a chi-square test to compare the fee and fine treatments ($\chi^2(1) = 0.73, p = 0.3933$), revealing no significant differences between them.

Regressions (6) also reveal differences across the individuals selected by the fee and the fine, exemplified by the *ControlDiff*, with the regular individual in the fine condition taking fewer points than the individual in the fee condition. This difference is not robust enough to be significant after controlling for the income effect in regression (7).

Result 4: *When participants take money in the treatment condition, both the* ***fee*** *and the* ***fine*** *lead to an increase in the amount taken compared to the control conditions.*

Result 4B: *There are no significant differences between the* ***fee*** *and* ***fine*** *in terms of the amount taken.*

In summary, our findings highlight the significant and heterogeneous impacts of introducing monetary penalties on prosocial behavior, with noteworthy distinctions between the fee and fine conditions. Some participants become less likely to take money after the penalty's

introduction, even if they had previously taken substantial amounts, indicating a crowding-in effect. Conversely, among participants who persist in taking money despite the penalty, they do so more intensively, demonstrating a crowding-out effect.

The fine condition effectively balanced these effects, resulting in no statistically significant impact on the overall amount of money taken. In contrast, the fee condition led to a substantial reduction, mainly due to significantly fewer instances of money being taken.

We also observed some differences in the impacts of different cases, and the relationship between inequality and behavioral changes is discussed and illustrated in Appendix D.

4 .2 Social Norms and Entitlement

In this section, we explore three potential mechanisms that may explain the observed behavioral changes: empirical expectations, normative expectations, and perceived entitlement.

We hypothesize a positive monotonic relationship between these behaviors and norms/entitlement, suggesting that if something is perceived as more expected, appropriate, or entitled, individuals are more likely to behave accordingly. To test this, we first examine whether the introduction of a monetary penalty affects these measures of social norms and entitlement. Then, we assess whether the observed behavioral changes can be linked to potential shifts in these measures, while controlling for treatment effects related to social norms/entitlement.

For each measure of social norms/entitlement, we assess two distinct aspects:

The first aspect reflects the extensive margin: For empirical expectations, we asked the participants to consider 100 other participants and inquire about how many would take money. For normative expectations, we inquired about the appropriateness levels of taking any amount, and for perceived entitlement, we asked how entitled the participant felt to take any amount. The regressions are illustrated in Table 6, with regressions (8)-(9)-(10) describing a linear regression with random effects for the empirical expectations, normative expectations, and entitlement, respectively:

	(8) Empirical	(9) Normative	(10) Entitlement
Fine	-5.866***	-1.812***	-1.010*
	(1.239)	(0.526)	(0.601)
Fee	-4.860***	-2.010***	-1.550***
	(1.484)	(0.477)	(0.442)
ControlDiff	3.476	-0.919	-0.693
	(2.432)	(0.838)	(0.962)
Constant	65.95***	34.30***	32.50***
	(1.645)	(0.635)	(0.721)
N	804	804	804

Standard errors clustered at the individual level in parentheses

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

Table 6: Social norms for the extensive margin: (8) empirical expectations, (9) normative expectations, and (10) perceived entitlement

For both the fee and the fine, participants expected fewer people to take money, perceived taking any amount of money as less socially appropriate, and attributed a lower perceived entitlement to take any amount of money. No significant difference between the fee and fine is observed.

Result 5 - Extensive Margin: *Both the fee and the fine lead to significant shifts in social norms associated with the extensive margin. Participants expect fewer people to take money and assign lower scores to normative and entitlement levels for taking any money.*

The second aspect is associated with the intensive margin. For empirical expectations, we initially inquired about the average amount of money taken by the same 100 participants. To better proxy the intensive margin, we weighted the expected amount by the expected number of participants taking money in each case, yielding the intensive margin. For normative and entitlement aspects, we asked participants to express the appropriateness/perceived entitlement for taking approximately 70% of the total amount.

The regressions are presented in Table 7. Regression (11) outlines a linear regression with random effects for empirical expectations, while regressions (12) depict the weighted empirical expectations. Regressions (13) and (14) present analyses for normative expectations and entitlement.

As we aim to examine the impact on the intensive margin and capture the crowding-out effect, we assess norm changes for those agents who continue taking money in the treatment condition. In other words, we analyze the norm change for the sample used in the previous intensive margin analysis.[8]

	(11) Empirical	(12) Weighted Empirical	(31) Normative	(14) Entitlement
Fine	7.947	142.3**	0.116*	-0.00265
	(7.361)	(62.64)	(0.0690)	(0.0814)
Fee	8.661	220.7	0.176**	0.128**
	(8.745)	(177.8)	(0.0699)	(0.0616)
ControlDiff	-23.45	-39.31	-0.155	0.0120
	(20.29)	(38.89)	(0.136)	(0.151)
Constant	365.2***	479.8***	3.083***	2.956***
	(14.59)	(33.71)	(0.0962)	(0.114)
N	556	546	556	556

Standard errors clustered at the individual level in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.001$

Table 7: Social norms for the intensive margin: (10) empirical expectations, (11) weighted empirical expectations, (12) normative expectations, and (13) perceived entitlement

The penalties do not induce changes in general empirical expectations regarding the amount taken. However, weighted empirical expectations display a significant increase in the

[8]For the Weighted Empirical Expectations, in a few cases, participants anticipated that no one would take money, preventing the creation of its weighted version.

fine condition and a substantial increase in the fee condition, though not deemed significant due to the high variance associated with this new measure.

Furthermore, both the fee and fine conditions lead to (marginally) significant increases in perceived appropriateness levels for taking larger sums of money. The fee condition also results in an increase in perceived entitlement to take larger amounts of money, while it has no effect on the fine condition.

Result 5 - Intensive Margin: *Both the fee and fine lead to some significant shifts in social norms related to the intensive margin. Participants assign higher scores to normative levels for taking any money in both fee and fine conditions. The fee leads to higher entitlement scores.*

To conclude our analysis, we incorporate social norms and entitlement into similar regression models as in the previous sections to investigate whether changes in social norms/entitlement could potentially explain behavioral changes.

We utilize behavioral observations from the four cases where we have measured social norms/entitlements to replicate the earlier findings. Then, we perform two new regressions: one to examine the relationship between social norms/entitlement and decisions, and another to examine the new treatment effects after adding social norms/entitlement.

Specifically, we initially run the same regression as before for the four observations (twin cases 2 and 4). As we are also examining the intensive and extensive margins, the extensive margin uses *Participation* as the dependent variable, as described in the regressions below:

$$Take_{i,r} = \beta_0 + \beta_1 Fine + \beta_2 Fee + \beta_3 ControlDiff + \epsilon_{i,r}$$

Subsequently, we conduct the following regressions:

$$Take_{i,r} = \hat{\beta}_0 + \beta_4 Empi + \beta_5 Norm + \beta_6 Enti + \epsilon_{i,r}$$

Where *Empi*, *Norm*, and *Enti* represent empirical expectations, normative expectations, and entitlement, respectively.

If β_4, β_5, and β_6 are significantly positive, the regression indicates a positive relationship between actions and behavior. For instance, if people consider larger amounts to be more socially appropriate, they are also more likely to participate.

Lastly, we use the following regression:

$$Take_{i,r} = \hat{\beta}_0 + \hat{\beta}_1 Fine + \hat{\beta}_2 Fee + \hat{\beta}_3 ControlDiff + \beta_4 Empi + \beta_5 Norm + \beta_6 Enti + \epsilon_{i,r}$$

With this specification, we test whether the treatment condition affects the amount taken through social norms. We can examine whether $\beta_1 = \hat{\beta}_1$ and $\beta_2 = \hat{\beta}_2$. If these coefficients are significantly different, it suggests that the treatment effects are influenced by variations in social norms between the treatment and control conditions, implying that changes in norms may partially explain the crowding-out (in) effects. Finally, we can test whether $\beta_1 - \beta_2 = \hat{\beta}_1 - \hat{\beta}_2$, which would indicate that the difference between the fee and fine treatments is influenced by changes in social norms across the conditions.

This model represents a mediation model, similar to those suggested by Howell (1992) and others. The general idea is that changes in social norms are correlated with changes in behavior, hence partially capturing the treatment effects. Here, we assume that the impact

of social norms and entitlement is consistent across the fee and fine conditions. In Appendix 15, we describe the robustness of these results, with only differences in the robustness of the entitlement.

In Table 8, regression (15) aims to replicate the previous results for the extensive margin using a smaller selected sample (2 twin cases in which the norms were measured) through linear[9] regression. In regression (16), we examine the relationship between social norms/entitlement and participation. In regression (17), we control for the treatment effects with the inclusion of social norms, checking if norms can partially explain the treatment effects. Regressions (18), (19), and (20) reproduce the same results for the intensive margin (Take) using linear regression.

	(15) Participation	(16) Participation	(17) Participation	(18) Take	(19) Take	(20) Take
Fine	-0.0644**		-0.0194	11.13		3.919
	(0.0250)		(0.0249)	(9.652)		(10.41)
Fee	-0.180***		-0.137***	24.02**		13.41
	(0.0280)		(0.0290)	(9.914)		(10.55)
ControlDiff	-0.00312		-0.0112	-24.30		-4.671
	(0.0343)		(0.0317)	(20.68)		(20.03)
Empirical		0.00501***	0.00488***		0.684***	0.684***
		(0.000661)	(0.000662)		(0.0434)	(0.0439)
Normative		0.00718***	0.00712***		1.636**	1.564**
		(0.00194)	(0.00195)		(0.703)	(0.718)
Entitlement		0.00329*	0.00339**		1.516**	1.516**
		(0.00172)	(0.00170)		(0.663)	(0.672)
Constant	0.815***	0.0866*	0.139**	395.4***	47.55**	47.91**
	(0.0242)	(0.0519)	(0.0539)	(15.94)	(18.67)	(20.80)
N	804	804	804	556	556	556

Standard errors in parentheses

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

Table 8: Channels: replicate extensive margin regressions (15) and norms controls (17), relation between norms and participation (16). Replicates intensive margin regressions (18) and norms controls (20), and relations between norms and amount taken (19).

First, regressions (15) and (18) almost perfectly replicate the results of regressions (4) and (7). The only difference lies in the significance of the fine treatment effect for the intensive margin, although it maintains the same directional value. This discrepancy could be partially explained by the fact that we utilize only half of the observations (those in which the norms were measured), and the results might be underpowered. However, all other results remain consistent across the regressions.

[9]To be able to compare the coefficients across the regressions.

Secondly, the coefficients for social norms and entitlement are positive and significant for all conditions and regressions. This indicates that measured social norms can partially explain behavioral levels. For example, if someone expects more people to take money, they are also more likely to take money. If someone thinks that it is more socially appropriate to take larger amounts of money, they will take more money.

Thirdly, we can check if the coefficients for the treatment effects and their differences change across the regressions:

For the extensive margin, the coefficients associated with the fine treatment effect are significantly different ($\chi^2(1) = 22.57, p = 0.000$), as are those for the fee condition ($\chi^2(1) = 29.44, p = 0.000$). However, the differences between the fee and fine conditions were not significantly explained by changes in social norms and entitlement ($\chi^2(1) = 0.00, p = 0.9810$). These results indicate that the decrease in coefficients for both fee and fine conditions is significant, and social norms partially account for the treatment effects. However, the gap between fee and fine conditions remains similar even when controlling for social norms.

For the intensive margin, the coefficients associated with the fine treatment effect are not significantly different ($\chi^2(1) = 1.64, p = 0.2000$). However, this result might partially be attributed to the fact that the coefficient itself was not significant in the replication (regression 17), leaving less room for the influence of social norms. Regarding the Fee condition, the coefficient change is marginally significant ($\chi^2(1) = 3.04, p = 0.0812$). Again, the difference between fee and fine was not significantly explained by changes in social norms and entitlement ($\chi^2(1) = 0.26, p = 0.6084$). These results indicate that the drop in coefficients for the fee condition is significant, while the decrease for the fine condition is illustrative but not statistically significant. Hence, social norms partially explain the treatment effects, especially for the fee condition.

Result 6: There is a positive correlation between the amount taken/participation and social norms/entitlement. The changes in social norms/entitlement partially account for the changes in the extensive and intensive margins in the fee condition.

Result 6B: Social norms/entitlement were unable to explain the differences between the fee and fine conditions.

The results indicate that the introduction of the fee and fine affects social norms and perceived entitlement. People expect fewer individuals to take money, find it less socially appropriate, and feel less entitled to take money. However, they also perceive taking larger amounts of money as more socially appropriate, and, in the fee condition, they also report higher levels of entitlement to take larger amounts of money.

These measures are positively correlated with behavior on both the extensive and intensive margins. For instance, if someone believes that more people take money or that it is more socially acceptable, they are more likely to take money themselves. Social norms and entitlements were able to partially capture the effects on both the intensive and extensive margins and can partially explain the crowding-out (in) effects. However, the changes in social norms and entitlement did not directly account for the differences between the treatment conditions (fee vs. fine).

Lastly, we can observe the distribution of changes in social norms and compare it with Figure 1, which shows the distribution of changes in the amount taken. Figure 5 illustrates the distribution of changes in empirical and normative expectations, grouping both the fee and fine:

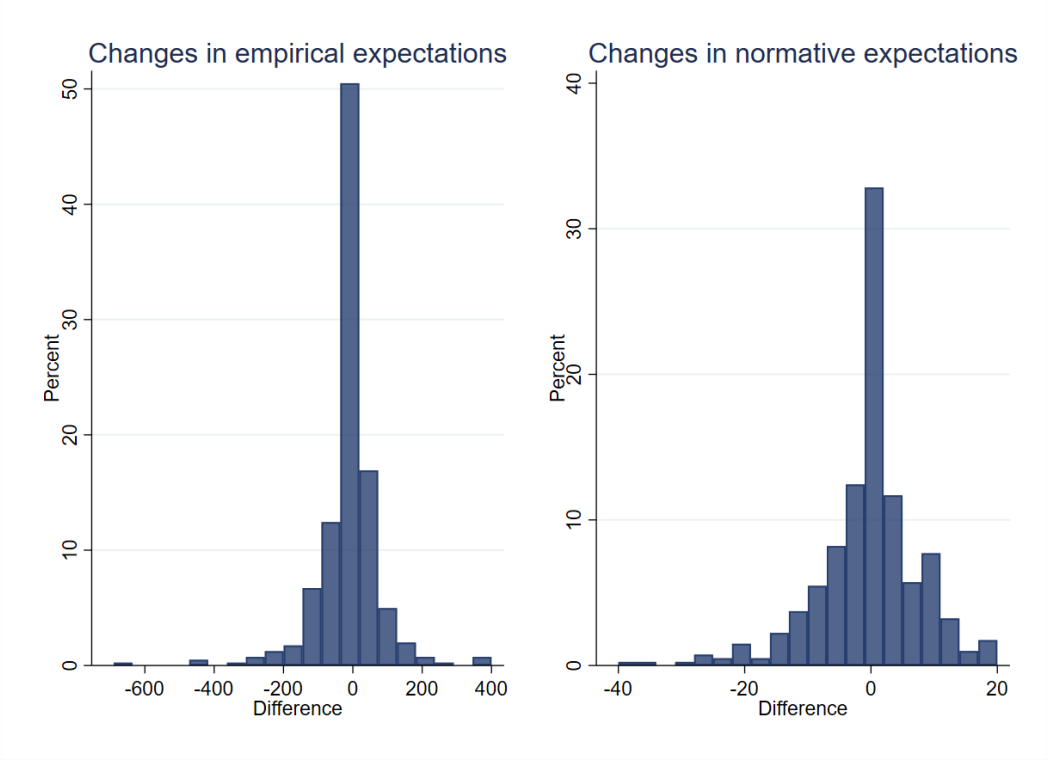

Figure 5: Left - Distribution of changes (treatment - control) for the empirical expectations, Right - Distribution of changes (treatment - control) for normative expectations

Once again, we observe that a significant number of participants report consistent norms for both conditions, mirroring what is observed for behavioral changes. Meanwhile, some individuals report significant increases and decreases in social norms.

5 Discussion

We investigate the impact of monetary penalties on prosocial motivation within a dictator game, comparing stylized fine and fee versions. Our goal is to address the theoretical contradictions that encompass both crowding-out (e.g., Gneezy and Rustichini (2000a))—leading to reduced prosocial motivation— and crowding-in effects (e.g., Kimbrough and Vostroknutov (2016))— which may boost prosocial motivation. This study aims to provide a nuanced understanding of the effects of monetary penalties on human behavior and offers insights into policy considerations that highlight how different penalty formats can lead to divergent outcomes. Lastly, we also investigate possible explanations for such behavioral changes.

While monetary penalties are commonly used to deter undesirable behaviors, their actual impact remains uncertain. Some studies demonstrate crowding-out effects (e.g., Gneezy and Rustichini (2000a); Frey and Jegen (2001); Gneezy et al. (2011)), suggesting that penalties may undermine individuals' prosocial concerns. Conversely, Kimbrough and Vostroknutov (2016, 2018), and others argue that individuals tend to comply with rules, even at the cost

of personal gain, when penalties are involved, leading to potential crowding-in effects. These conflicting findings raise questions about the net impact of penalties.

We find that monetary penalties have heterogeneous impacts on different individuals. Our observations reveal both crowding-in and crowding-out effects, providing explanations for the apparent contradictions.

First, a substantial number of participants consistently choose the same monetary amounts in both control and treatment conditions. These participants could be following models associated with prosocial preferences (e.g., Fehr and Schmidt (1999); Andreoni and Miller (2002)), as they would predict no changes. However, we also observe that these participants maintain consistent social norms across conditions, aligning with models proposed by Krupka and Weber (2013) and Akerlof and Kranton (2000). This indicates that for some individuals, the treatment effects have minimal influence on their norms and subsequently on their decisions, which aligns with our following discussion on social norms.

Secondly, when the penalty is introduced, many participants refrain from taking money, even if they had previously taken large amounts. The fee treatment leads to a roughly 15% reduction in the number of cases where money is taken, while the fine treatment results in a 5% reduction. Notably, the majority of those participants were taking substantial amounts of money, with about 50% taking more than 200 points, and approximately 15% were taking all the money available. These behavioral changes suggest a rule-following tendency similar to Kimbrough and Vostroknutov (2016) and can be seen as a crowding-in effect.

Thirdly, among participants who continue taking money after the penalty's implementation, there is a significant increase in the amount they take. This demonstrates that the penalty motivates these individuals to act more selfishly and take money more intensively, indicating a crowding-out effect.

These crowding-in and crowding-out effects represent opposing forces, and the context influences these forces, determining which effect prevails at the aggregate level. This observation is evident in our experiment when analyzing the differences between fees and fines.

We designed fees and fines to be as natural and directly comparable as possible. Fees impose penalties before the action, creating a first-stage decision where participants must decide whether to take money. Fines are subtracted after taking any money.

At an aggregate level, fines appear inefficient and do not significantly change the average amount taken in the experiment. This does not imply that fines do not impact behavior; instead, it suggests that crowding-in and crowding-out effects balance each other out, resulting in a null aggregate effect. In fact, participants who continued taking money increased their amount taken by almost 4%, offsetting the 5% decrease in cases where money ceased to be taken due to the fine.

Conversely, the fee condition has a significant impact, leading to a nearly 9% reduction in the average amount of money taken. This difference is reflected in the roughly 15% decrease in the number of cases where money is taken, while participants who continue taking money increase the amount taken by around 6%.

The disparities between the fee and fine conditions are primarily driven by the extensive margin. The fee results in significantly fewer people taking money compared to the fine, while changes at the intensive margin are not significantly different. Since the participants across conditions are similar, this further drop reflects a stronger crowding-in effect and a greater tendency to follow the rules in the fee condition compared to the fine condition. In other words, participants act more prosocially in the fee condition than in the fine condition, even though both involve the same fixed cost.

In conclusion, our results reveal the heterogeneous impacts of monetary penalties, with

some participants being consistent, some following rules and experiencing crowding-in effects, while others seize the opportunity to take more money, resulting in crowding-out effects. The balance of these forces depends on the context, with the fee condition proving more efficient than the fine since it effectively reduces the aggregate amount taken and exhibits a stronger crowding-in effect.

Our findings also open avenues for future research. Firstly, investigating the relationship between penalty severity and its influence on crowding-in and crowding-out effects can provide valuable insights. Higher penalty levels may reduce the frequency of occurrences but also potentially amplify crowding-out effects. Further research could explore the identification of individuals more likely to exhibit crowding-out or crowding-in behavior in response to monetary incentives, allowing for more tailored penalty designs.

Additionally, understanding the nuanced implications of interventions and policy behavioral designs, as proposed by Bowles (2016), warrants exploration. This extends to areas like environmental legislation employing fines to deter environmental damage alongside the establishment of licenses (similar to fees) to permit specific behaviors. Contemporary approaches, such as carbon markets, might bring about even more moral changes, as described by Falk and Szech (2013). These different setups may appear similar in economic theory but can elicit distinct behavioral responses and trigger different moral perceptions. Contrasting policy tools in more realistic settings is essential for designing more effective policies and interventions.

Our study also delves into potential mechanisms behind behavioral changes. First, we design our experiment to minimize the impact of other potential explanations. We use a dictator game, which minimizes strategic interaction and incomplete information—factors often responsible for crowding-out effects (e.g., Bénabou and Tirole (2006) and Frey and Jegen (2001))—and also reduces the influence of other potential factors that could drive behavioral changes.

We investigate the roles of social norms and perceived entitlement as potential explanations for behavioral changes. Regarding social norms, we categorize them into empirical and normative expectations, following the terminology of Bicchieri (2016, 2005). Social norms and entitlement have been identified as plausible mechanisms behind the crowding-out (or crowding-in) effects, as explained by Capraro and Perc (2021); Ellingsen and Mohlin (2022); Frey and Jegen (2001); Bénabou and Tirole (2006); Gneezy et al. (2011). Our experimental setup provides an opportunity to directly test this theory while minimizing the influence of other possible factors.

To measure entitlement, we adapted the methodology from Krupka and Weber (2013), employing a coordination game to incentivize participants to consider the group's opinion. This approach aligns with attribution theory in social psychology, as discussed in Peterson et al. (1982) and Dykema et al. (1996), which explores how individuals perceive the causes and motivations behind everyday experiences.

We encourage further research to expand these methods to capture additional aspects related to concerns about social image and motivated reasoning, as explored in previous studies, such as Epley and Gilovich (2016). For example, Fischer and Teixeira (2023) employs a similar methodology to examine how different motivations are attributed to males or females for the same behavior, shedding light on gender differences and their underlying causes.

Our study demonstrates that the implementation of monetary penalties shifts the social norms. Participants perceive, for example, that others are less likely to take money when penalties are in place and that taking larger amounts of money is more socially acceptable.

Intuitively, the logic is: "You should not do it, if you do, you should make the most of it".

These results corroborate with Lane et al. (2023a), which describes how the implementation of a law can trigger different social norms. This is consistent with findings by Ellingsen et al. (2012) and Eriksson et al. (2017), which show that framing effects can influence behavior and expectations. In other words, incentives can alter how individuals perceive a situation, not just the trade-offs associated with it.

Our research also establishes a positive correlation between norms and behavior, both at the extensive and intensive margins. This implies that individuals who believe that taking more money is socially appropriate are more likely to do so, highlighting their conformity to social norms. This aligns with the extensive literature (e.g., Kimbrough and Vostroknutov (2016); Bicchieri (2005); Krupka and Weber (2013); Xiao and Bicchieri (2010)) that demonstrates the relationship between social norms and behavior.

Moreover, changes in social norms partially account for the observed behavioral shifts, mediating the treatment effects. We use a regression model to demonstrate that the treatment effects associated with the fee and fine are partially explained by the changes in social norms. This suggests that alterations in norms can, to some extent, explain crowding-in and crowding-out effects.

These findings provide direct evidence that crowding-out and crowding-in effects can be partially attributed to social norms. This is particularly significant in the context of a dictator game, emphasizing that social concerns influence individual decision-making, and minimize the strategical interaction aspects. Hence, our study reveals that crowding-in and crowding-out effects can be observed in very simple settings.

These results have substantial implications for policy interventions, aligning with discussions in works such as those by Bicchieri and Dimant (2019), Sunstein (2003), and Lane, Nosenzo, and Sonderegger (2023b). They underscore the need for careful execution of interventions to avoid unintended crowding-out effects. Moreover, they highlight the intricate interplay between context and norms and the potential for indirect interventions through context design. In other words, rather than directly targeting behavior or social norms, a possible intervention might attempt to use framing effects to change the perception of the situation, indirectly shaping norms.

While social norms can partially explain the crowding-in and crowding-out effects, they are insufficient to elucidate the distinctions between the fee and fine conditions in our setting. Other factors integrated into our experimental design could also contribute to the differences between the fee and fine conditions. For instance, the first-stage decision in the fee condition may induce narrow bracketing (e.g., Read et al. (2000)) by isolating the problem from the broader context, leading to a different cognitive process. Another possibility is related to Zellermayer (1996), as the first-stage decision may make the payment more salient, leading to stronger emotional responses. Such cognitive and emotional responses might trigger behavioral changes without significantly impacting the observed social norms. Future research may seek to further dissect these differences, which can be crucial for designing better and more precise interventions.

Our study not only contributes to a better understanding of the impact of monetary penalties on behavior but also delves into the heterogeneous effects these penalties can have. By thoroughly analyzing the motivations behind these behavioral changes, we pave the way for a more comprehensive comprehension of how financial incentives and deterrents shape individual decision-making. These findings generate valuable insights for future research and have the potential to inform the design of more effective policy interventions.

6 Conclusion

Monetary penalties are a common tool for discouraging undesirable behavior, yet their precise impact remains uncertain, and divergent behavior has been observed. Our findings shed light on these contradictory theories by revealing the heterogeneous effects of monetary penalties on individuals. For many, the penalty serves as motivation to cease undesirable behavior, even when they were previously engaging in it intensely, illustrating crowding-in effects. On the other hand, some individuals intensify their undesirable behavior, demonstrating a crowding-out effect.

The interplay of these forces is context-dependent, as exemplified by the different impacts observed with fees and fines. In our study, fines exhibit a balance of these heterogeneous effects and produce no significant aggregate impact. However, when the same penalty is introduced as a fee, paid before the action, it proves effective, with the crowding-in effect dominating the crowding-out effect, resulting in dictators taking less on average from receivers and fewer dictators taking money. This underscores how subtle contextual changes can influence the efficiency of penalties.

Furthermore, our observations indicate that the introduction of monetary penalties shifts perceived social norms. People expect fewer individuals to engage in the target behavior when a penalty is in place, while also believing that engaging in the behavior intensively is more socially acceptable with the penalty compared to when it is absent. These shifts in social norms can partially account for behavioral changes, providing a partial explanation for both crowding-in and crowding-out effects.

In summary, we demonstrate that monetary penalties have heterogeneous impacts on behavior, leading to crowding-out effects and crowding-in effects for different individuals. The format of the penalty affects these effects, with significant differences depending on whether the penalty is paid after (as in a fine) or before (as in a fee). The implementation of monetary penalties not only affects the trade-off but also shifts the perceived norms associated with the situation. These changes in norms can partially explain the crowding-out and crowding-in effects.

References

Akerlof, G. A., & Kranton, R. E. (2000). Economics and identity. *The Quarterly Journal of Economics*, *115*(3), 715–753.

Andreoni, J., & Miller, J. (2002). Giving according to garp: An experimental test of the consistency of preferences for altruism. *Econometrica*, *70*(2), 737–753.

Becker, G. S. (1968). Crime and punishment: An economic approach. *Journal of political economy*, *76*(2), 169–217.

Bicchieri, C. (2005). *The grammar of society: The nature and dynamics of social norms.* Cambridge University Press.

Bicchieri, C. (2016). *Norms in the wild: How to diagnose, measure, and change social norms.* Oxford University Press.

Bicchieri, C., & Dimant, E. (2019). Nudging with care: The risks and benefits of social information. *Public Choice*, 1–22.

Bolton, G. E., & Ockenfels, A. (2000). Erc: A theory of equity, reciprocity, and competition. *American Economic Review*, *91*(1), 166–193.

Bordalo, P., Gennaioli, N., & Shleifer, A. (2013). Salience and consumer choice. *Journal of Political Economy*, *121*(5), 803–843.

Bowles, S. (2016). *The moral economy: Why good incentives are no substitute for good citizens.* Yale University Press.

Bénabou, R., & Tirole, J. (2003). Intrinsic and extrinsic motivation. *The Review of Economic Studies*, *70*(3), 489–520.

Bénabou, R., & Tirole, J. (2006). Incentives and prosocial behavior. *American Economic Review*, *96*(5), 1652–1678.

Capraro, V., & Perc, M. (2021). Mathematical foundations of moral preferences. *Journal of the Royal Society Interface*, *18*(175), 20200880.

Charness, G., & Rabin, M. (2002). Understanding social preferences with simple tests. *The Quarterly Journal of Economics*, *117*(3), 817–869.

Chen, D. L., Schonger, M., & Wickens, C. (2016). otree—an open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance*, *9*, 88–97.

Cialdini, R. B., & Trost, M. R. (1998). Social influence: Social norms, conformity and compliance.

Dykema, J., Bergbower, K., Doctora, J. D., & Peterson, C. (1996). An attributional style questionnaire for general use. *Journal of Psychoeducational Assessment*, *14*(2), 100–108.

Ellingsen, T., Johannesson, M., Mollerstrom, J., & Munkhammar, S. (2012). Social framing effects: Preferences or beliefs? *Games and Economic Behavior*, *76*(1), 117–130.

Ellingsen, T., & Mohlin, E. (2022). *A model of social duties.*

Epley, N., & Gilovich, T. (2016). The mechanics of motivated reasoning. *Journal of Economic perspectives*, *30*(3), 133–140.

Eriksson, K., Strimling, P., Andersson, P. A., & Lindholm, T. (2017). Costly punishment in the ultimatum game evokes moral concern, in particular when framed as payoff reduction. *Journal of Experimental Social Psychology*, *69*, 59–64.

Falk, A., Becker, A., Dohmen, T., Enke, B., Huffman, D., & Sunde, U. (2018). Global evidence on economic preferences. *The Quarterly Journal of Economics*, *133*(4), 1645–1692.

Falk, A., & Szech, N. (2013). Morals and markets. *Science*, *340*(6133), 707–711.

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

Fischer, P., & Teixeira, R. (2023). *Sex, lies, and punishment: Gender differences in receiving punishment after suspected dishonesty.*

Frey, B. S. (2000). Not just for the money: An economic theory of motivation. *Financial Counseling and Planning*, *11*(1).

Frey, B. S., & Jegen, R. (2001). Motivation crowding theory. *Journal of Economic Surveys*, *15*(5), 589–611.

Frey, B. S., & Oberholzer-Gee, F. (1997). The cost of price incentives: An empirical analysis of motivation crowding-out. *The American economic review*, *87*(4), 746–755.

Gneezy, U., Meier, S., & Rey-Biel, P. (2011). When and why incentives (don't) work to modify behavior. *Journal of Economic Perspectives*, *25*(4), 191–210.

Gneezy, U., & Rustichini, A. (2000a). A fine is a price. *The Journal of Legal Studies*, *29*(1), 1–17.

Gneezy, U., & Rustichini, A. (2000b). Pay enough or don't pay at all. *The Quarterly Journal of Economics*, *115*(3), 791–810.

Hong, S.-M., & Faedda, S. (1996). Refinement of the hong psychological reactance scale. *Educational and Psychological Measurement*, *56*(1), 173–182.

Howell, D. C. (1992). *Statistical methods for psychology.* PWS-Kent Publishing Co.

Janssen, M. C., & Mendys-Kamphorst, E. (2004). The price of a price: On the crowding out and in of social norms. *Journal of Economic Behavior Organization*, *55*(3), 377–395.

Kimbrough, E. O., & Vostroknutov, A. (2016). Norms make preferences social. *Journal of the European Economic Association*, *14*(3), 608–638.

Kimbrough, E. O., & Vostroknutov, A. (2018). A portable method of eliciting respect for social norms. *Economics Letters*, *168*, 147–150.

Kornhauser, L., Lu, Y., & Tontrup, S. (2020). Testing a fine is a price in the lab. *International Review of Law and Economics*, *63*, 105931.

Krupka, E. L., & Weber, R. A. (2013). Identifying social norms using coordination games: Why does dictator game sharing vary? *Journal of the European Economic Association*, *11*(3), 495–524.

Kurz, T., Thomas, W. E., & Fonseca, M. A. (2014). A fine is a more effective financial deterrent when framed retributively and extracted publicly. *Journal of Experimental Social Psychology*, *54*, 170–177.

Lane, T., Nosenzo, D., & Sonderegger, S. (2023a). Law and norms: Empirical evidence. *American Economic Review*, *113*(5), 1255–1293.

Lane, T., Nosenzo, D., & Sonderegger, S. (2023b). Law and norms: Empirical evidence. *American Economic Review*, *113*(5), 1255–1293.

Mellström, C., & Johannesson, M. (2008). Crowding out in blood donation: Was titmuss right? *Journal of the European Economic Association*, *6*(4), 845–863.

Peterson, C., Semmel, A., Von Baeyer, C., Abramson, L. Y., Metalsky, G. I., & Seligman, M. E. (1982). The attributional style questionnaire. *Cognitive therapy and research*, *6*(3), 287–299.

Read, D., Loewenstein, G., Rabin, M., Keren, G., & Laibson, D. (2000). Choice bracketing. *Elicitation of preferences*, 171–202.

Sunstein, C. R. (2003). Moral heuristics and moral framing. *Minnesota Law Review*, *88*, 1556.

Titmuss, R. M., et al. (1970). *The gift relationship.* Allen & Unwin London.

Tonin, M., & Vlassopoulos, M. (2013). Experimental evidence of self-image concerns as motivation for giving. *Journal of Economic Behavior & Organization*, *90*, 19–27.

Xiao, E., & Bicchieri, C. (2010). When equality trumps reciprocity. *Journal of Economic Psychology*, *31*(3), 456–470.

Yang, Y., Onderstal, S., & Schram, A. (2016). Inequity aversion revisited. *Journal of Economic Psychology*, *54*, 1–16.

Zellermayer, O. (1996). *The pain of paying*. Carnegie Mellon University.

Appendix

A Quadratic inequality aversion

The utility function, denoted as U, encapsulates the agent's concern for their initial endowment (x), the amount they decide to take (t), and introduces a negative weighting factor, $\beta > 0$, to express the quadratic relationship between their gains and the gains of others, expressed as $((x+t)-(y-t))^2$.

In the treatment condition, applicable to both the fee and fine scenarios, an additional penalty of 100 points is incurred if the agent chooses to take points. This leads to the following optimization problem as shown below:

$$\max_t : U(t) = \begin{cases} x - \beta(x-y)^2 & \text{if } t = 0 \\ x + t - 100 - \beta(x - y - 100 + 2t)^2 & \text{if } t > 0 \end{cases}$$

By solving the optimization problem for the case in which $t > 0$, we deduce that the maximum argument is $t = \frac{1}{8}(400 + \frac{1}{\beta} - 4x + 4y)$, and the maximum value is $\frac{1+8\beta(-100+x+y)}{16\beta}$. The agent will take zero if:

$$x - \beta(x-y)^2 > \frac{1 + 8\beta(-100 + x + y)}{16\beta}$$

Notice that each case creates a different initial inequality, which the agent will maintain if the agent does not take money. As for all cases $(-100 + x + y) = 900$, we can simplify the problem into:

$$x - \beta(x-y)^2 > 450 + \frac{1}{16\beta}$$

We can analyze this inequality for all cases in our experiment and check the β in which the agent will stop taking money for each condition. By resolving this inequality for all conceivable scenarios[10], the possible solutions for the inequality are[11]:

$$x = 600, y = 400, \frac{3-\sqrt{5}}{1600} < \beta < \frac{3+\sqrt{5}}{1600}$$

$$x = 650, y = 350, \frac{4-\sqrt{7}}{3600} < \beta < \frac{4+\sqrt{7}}{3600}$$

Now, we can check how much money such a participant was taking in the control conditions, given the β values and their respective cases:

$$x = 500, y = 400, 0 < t \leq 80.90$$

$$x = 550, y = 350, 0 < t \leq 66.14$$

Hence, the maximum amount that the dictator would take before stopping would be 80.90.

[10] Notice that we are analyzing only the twin cases, but a similar argument could be made for all cases.

[11] If the agent starts behind, there is no solution with positive β

B Balance table

The Table 9 describes the demographics across conditions (using the Profilic data):

	(Fine)	(Fee)	(Difference)
	Mean/SD	Mean/SD	Difference/p-value
Time	1130.76	1287.37	-156.61*
	(400.53)	(577.68)	[0.03]
Age	39.43	39.75	-0.32
	(12.84)	(11.98)	[0.86]
Gender	0.50	0.43	0.07
	(0.50)	(0.50)	[0.32]
Ethnicity	0.84	0.82	0.02
	(0.37)	(0.39)	[0.73]
Observations	100	100	200

* p<0.05, ** p<0.01, *** p<0.001
Standard deviation in parentheses
t statistics in brackets

Table 9: Balance Table

Participants are similar between the fine and fee groups. However, people consistently take more time in the fee condition.

C Order Effects

TTable 10 provides an analysis of the amount taken by condition, comparing the order of the session. We use the following regression:

$$Take_{i,r} = \beta_0 + \beta_1 Session + \beta_2 Order + \beta_3 Session \times Order + \epsilon_{i,r}$$

Session is a dummy variable that is equal to 1 if the fee is applied in that specific session, *Order* is a dummy variable that is equal to 1 if the session started with the treatment condition. Lastly, there is an interaction term that evaluates whether the order effect may be different for the Fee or the Fine conditions.

	(Control) Take	(Treatments) Take
Session	5.667	-27.62**
	(11.97)	(13.08)
Order	2.177	-13.06
	(27.64)	(29.69)
Session × Order	-0.487	8.540
	(17.01)	(18.49)
Constant	276.9***	289.8***
	(19.45)	(21.00)
N	2020	2000

Robust standard errors in parentheses

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

Table 10: Regression (Control) describes order effects for the control conditions, Regression (Treatments) describes order effects for the treatment conditions

Regression (Control) illustrates the order effects on the control conditions, using the observations only associated with the control. Regression (Treatments) illustrates the order effects on the treatment conditions.

The results showed significant differences between the fee and fine treatments, while no impact on the order was observed.

D Cases & Inequality

D .1 Cases

We observe that the cases play a role in individuals' behavior. To simplify the discussion, we focus on the control conditions, avoiding the income effect associated with the treatment, and observe how the amount taken varies across different situations. We run the following regression:

$$Total_{i,r} = \beta_0 + \beta_i case_i + \epsilon_{i,r}$$

Total indicates the sum of the endowment with the amount taken, and we also use one dummy for each case. The results can be observed in Table 11:

	(1) Total	(2) Total	(3) Total	(4) Participation
170	10.75		10.75	2.14e-15
	(6.666)		(6.668)	(1.806)
200			68.91***	2.25e-15
			(7.297)	(1.806)
270		16.22**	85.12***	2.77e-15
		(7.875)	(7.823)	(1.806)
360.			7.910	1.66e-15
			(8.010)	(1.806)
500	47.91***		47.91***	-13.42***
	(8.537)		(8.541)	(1.995)
550	81.94***		81.94***	-12.59***
	(8.692)		(8.695)	(1.963)
600		84.63***	153.5***	-13.11***
		(8.933)	(9.273)	(1.983)
620			91.89***	-13.52***
			(9.772)	(1.998)
650		112.3***	181.2***	-12.70***
		(8.822)	(8.446)	(1.967)
Constant	609.7***	678.6***	609.7***	16.35***
	(13.04)	(13.81)	(13.05)	(2.060)
N	804	804	2010	2010

Standard errors in parentheses

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

Table 11: Regression (1) describes the impact of the cases in which the total sum is 900, Regression (2) for a total sum of 1000, Regression (3) includes all data, and Regression (4) checks the participants across conditions

Regression (4) shows that almost all participants take money when they are behind, and many stop taking money when they are ahead. The proportion of agents who cease is fairly consistent for all cases in which they are ahead.

Regressions (1-2-3) show that participants consistently keep a higher proportion of the total share when they have higher endowments.

To continue this analysis, we run the following regression:

$$Total_{i,r} = \beta_0 + \beta_1 Endowment + \beta_2 1000\text{-}Total + \epsilon_{i,r}$$

We analyze the total taken, considering a linear relation for the endowment, and we add a dummy to control if the case is dividing 1000 points or 900 points. The results can be

observed in Table 12:

	(1) Total	(2) Total	(3) Total
Endowment	0.193**	0.617***	0.196***
	(0.0766)	(0.0855)	(0.0161)
1000-Total	52.38***	40.72***	67.49***
	(8.744)	(8.945)	(3.681)
Constant	589.0***	350.4***	580.2***
	(16.43)	(49.63)	(13.89)
N	804	804	1608

Standard errors in parentheses

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

Table 12: Regression (1) describes the impact of endowment for cases in which the dictator starts behind, Regression (2) for cases in which the dictator starts ahead, and Regression (3) includes all data

When the agent is behind, an increase of one unit in endowment leads to a 0.20 increase in the total amount kept. When the agent is ahead, each unit increase leads to a 0.60 increase in the total amount kept.

Hence, the results indicate that agents have some reference dependence aspect associating endowments and the amount taken. Future research might aim to further understand these aspects of decision-making.

Please note that our results directly compare the same cases (twin cases), so this observed tendency does not directly affect the results presented in the main findings.

D .2 Inequality

We investigate whether the distribution of the initial endowment has an impact on the results observed in the main behavioral section. Specifically, we analyze whether the starting point of the dictators, either with more or fewer points than the receiver, influences the effectiveness of the monetary penalty in inducing behavioral change.

To do so, we will re-perform all the analyses and split the cases into two possibilities: dictators starting ahead or behind the participants. We will re-perform all the regressions, first using the subsample of each situation (ahead or behind), and then by adding an interaction term between treatments and inequality. Moreover, we will directly compare the twin cases, which control for income effects and serve as the main benchmark of our results.

We begin by analyzing the aggregate results, which can be observed in Table 13:

	(1) Take	(2) Take	(3) Take
ControlDiff	-4.030	-6.215	-5.123
	(24.77)	(20.16)	(21.58)
Fine	2.475	-14.80	3.019
	(10.06)	(9.195)	(10.44)
Fee	-4.750	-50.80***	-5.299
	(14.42)	(11.49)	(13.89)
Ahead			-330.4***
			(6.668)
Fine × Ahead			-18.36
			(11.82)
Fee × Ahead			-44.95***
			(15.87)
Constant	482.1***	152.7***	482.6***
	(18.50)	(14.47)	(17.17)
N	804	804	1608

Standard errors clustered at the individual level in parentheses

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

Table 13: Regression (1) describes the impact of treatment on the amount taken for cases in which the dictator starts behind, Regression (2) for cases in which the dictator starts ahead, and Regression (3) includes all data

The results reveal that the Fee condition is only effective when the agent is in a leading position.

When the agent is behind, both the fee and fine conditions lead to a reduction, but the significance of this reduction varies. Regression (6) shows a significant impact, whereas regression (4) does not demonstrate significance.

The results indicate that both the fee and fine conditions lead to a significant reduction when the agents are ahead. However, once again, the results are mixed. In the case of the Fine condition, regression (5) shows a significant impact, while regression (6) is not statistically significant.

The difference in the extensive margin between the fee and fine conditions is significantly more pronounced when the agent is ahead, and this difference is only significant in this situation.

Lastly, we analyze the intensive margin, and the results can be observed in Table 14:

	(7) Take	(8) Take	(9) Take
ControlDiff	-4.759	-28.31	-3.171
	(25.19)	(25.92)	(23.35)
Fine	11.62	22.69**	11.04
	(8.759)	(10.71)	(9.006)
Fee	22.38**	33.33**	22.99**
	(10.03)	(13.28)	(9.517)
Ahead			-331.4***
			(8.684)
Fine × Ahead			12.76
			(11.30)
Fee × Ahead			8.668
			(14.09)
Constant	484.0***	256.7***	483.2***
	(18.95)	(20.21)	(18.07)
N	772	346	1118

Standard errors clustered at the individual level in parentheses

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

Table 14: Regression (7) describes the impact of treatment on the amount taken for cases in which the dictator starts behind, Regression (8) for cases in which the dictator starts ahead, and Regression (9) includes all data

The results for the intensive margin show that the crowding-out effect is fairly consistent across situations. The fine condition leads to a nonsignificant increase when the agent is behind, while the fee condition is significant. Both conditions are significant when the agent is ahead, and regression (9) replicates these results.

In general, the results indicate that the crowding-out effect is fairly consistent whether the agent is ahead or behind, with some evidence that it can lead to slightly bigger impacts when the agent is ahead.

However, the rule-following tendency and potential crowding-in effects do not necessarily have the same partner. It was observed that the majority of the participants still take money when they are behind, and both the fee and fine lead to a reduction, though relatively smaller. When the agent is ahead, both the fee and fine seem to be effective, with the fee being even more effective.

The aggregate results follow the balance of these two forces, with no impacts when the agent is behind, and the fee being effective when the agent is ahead.

Future research might further explore these differences and seek to better understand the reasoning behind these behavioral channels.

Potentially, the agents face higher moral costs when the agent is ahead, leading to differences in the extensive margin. However, given that the agent is willing to take money,

the presence of the penalty leads to a decision to take more money.

E Norms and Entitlement: Robustness

The regression illustrated in Table 15 presents the same results as Table 8 and introduces an interaction factor between the norms and the fee condition. This analysis aims to determine whether norms exhibit different behaviors under fee and fine conditions. Therefore, we use the following equations:

$$Take/Participation_{i,r} = \beta_0 + \beta_1 Fine + \beta_2 Fee + \beta_3 ControlDiff + \epsilon_{i,r}$$

$$Take/Participation_{i,r} = \hat{\beta}_0 + \hat{\beta}_1 Fine + \hat{\beta}_2 Fee + \hat{\beta}_3 ControlDiff + \beta_4 Empi + \beta_5 Norm + \beta_6 Enti + \epsilon_{i,r}$$

$$\begin{aligned} Take/Participation_{i,r} = \hat{\beta}_0 + \hat{\beta}_1 Fine + \hat{\beta}_2 Fee + \hat{\beta}_3 ControlDiff + \beta_4 Empi + \beta_5 Norm + \beta_6 Enti \\ + \beta_7 Empi \times Fee + \beta_8 Norm \times Fee + \beta_9 Enti \times Fee + \epsilon_{i,r} \end{aligned}$$

The results can be observed in the table below:

	(1) Participation	(2) Participation	(3) Participation	(4) Take	(5) taTakeke	(6) Take
Fine	-0.0644**	-0.0194	-0.0258	11.13	3.919	4.713
	(0.0250)	(0.0249)	(0.0254)	(9.652)	(10.41)	(10.38)
Fee	-0.180***	-0.137***	-0.131***	24.02**	13.41	12.86
	(0.0280)	(0.0290)	(0.0298)	(9.914)	(10.55)	(10.91)
ControlDiff	-0.00312	-0.0112	0.122	-24.30	-4.671	33.51
	(0.0343)	(0.0317)	(0.107)	(20.68)	(20.03)	(38.84)
Empirical		0.00488***	0.00564***		0.684***	0.744***
		(0.000662)	(0.000857)		(0.0439)	(0.0658)
Normative		0.00712***	0.00941***		1.564**	2.124**
		(0.00195)	(0.00249)		(0.718)	(1.000)
Entitlement		0.00339**	0.00156		1.516**	0.775
		(0.00170)	(0.00205)		(0.672)	(0.921)
Empirical × Fee			-0.00157			-0.118
			(0.00129)			(0.0890)
Normative × Fee			-0.00418			-0.840
			(0.00369)			(1.444)
Entitlement × Fee			0.00359			0.986
			(0.00315)			(1.341)
Constant	0.815***	0.139**	0.0698	395.4***	47.91**	31.10
	(0.0242)	(0.0539)	(0.0642)	(15.94)	(20.80)	(27.61)
N	804	804	804	556	556	556

Standard errors clustered at the individual level in parentheses

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

Table 15: Channels: robustness check

The results are quite consistent, except for perceived entitlement.

F Instructions

Introduction, instructions, and example of comprehension check:

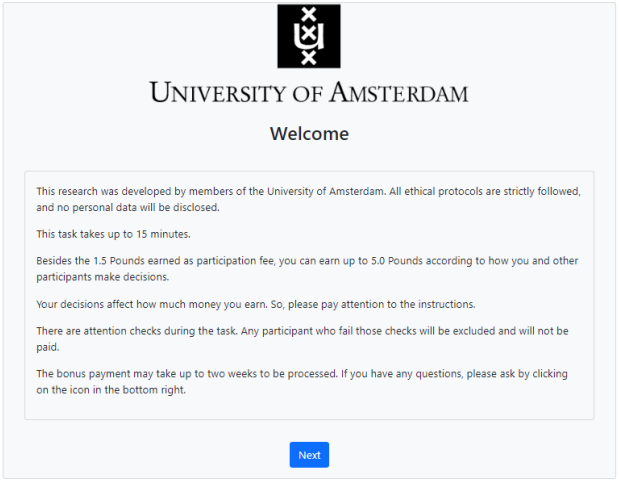

Figure 6: Introduction

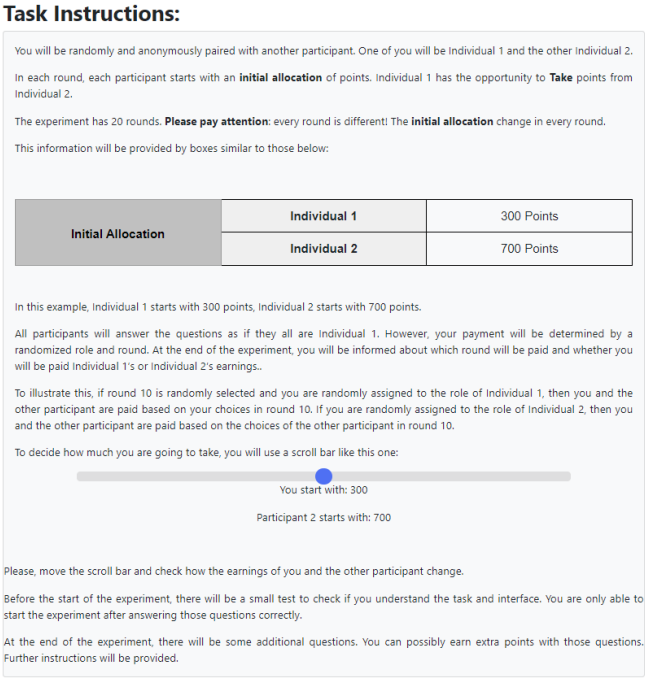

Figure 7: Instructions

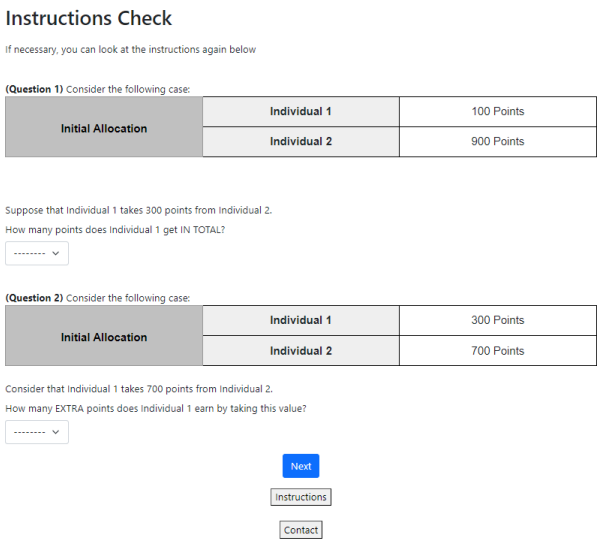

Figure 8: Example - Comprehension check

Decision - Control, info fine, fine, info fee, and fee:

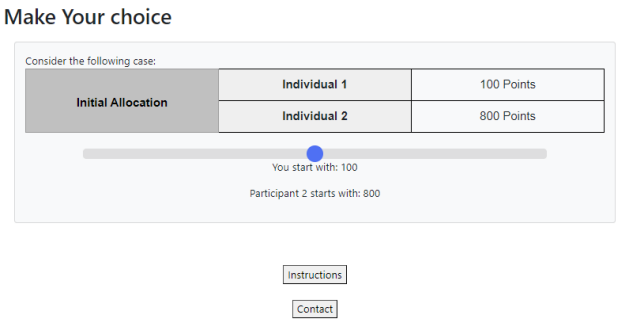

Figure 9: Example: Control Condition

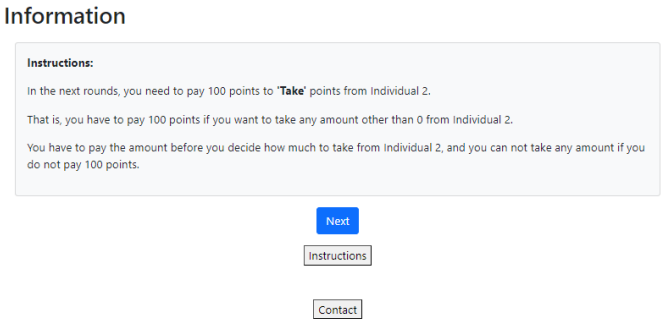

Figure 10: Information - Fine

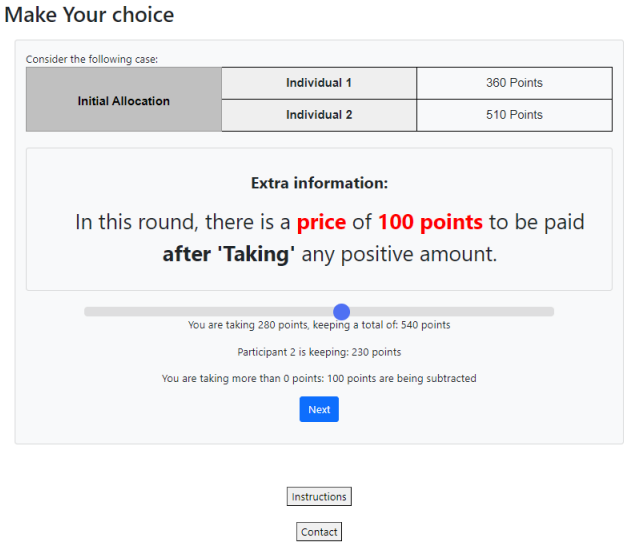

Figure 11: Example: Fine Condition

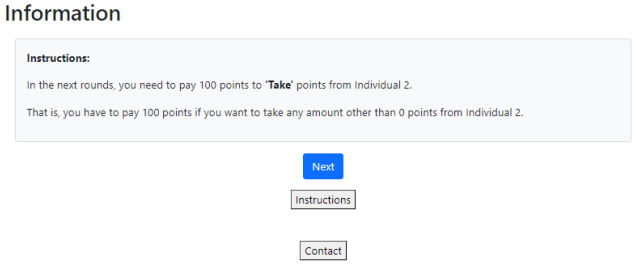

Figure 12: Information - Fee

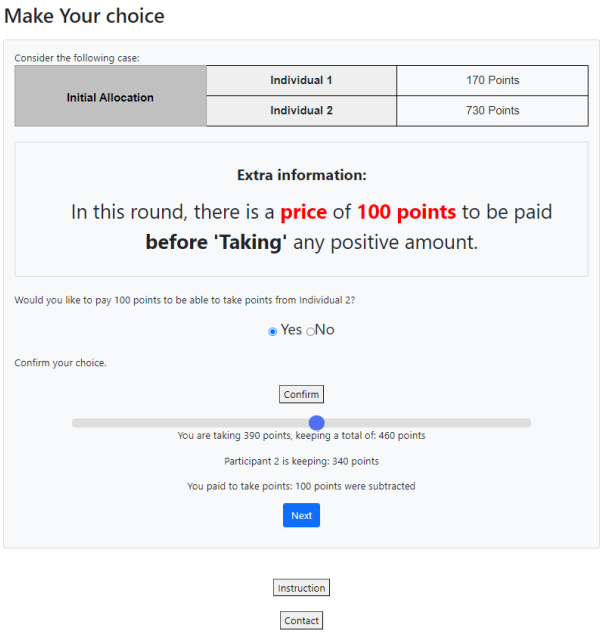

Figure 13: Example: Fee Condition

Social Norms and Entitlement:

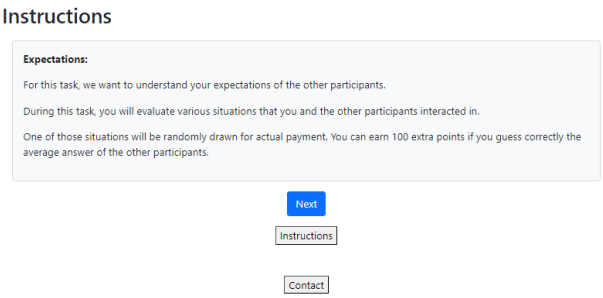

Figure 14: Information - Empirical Expectation

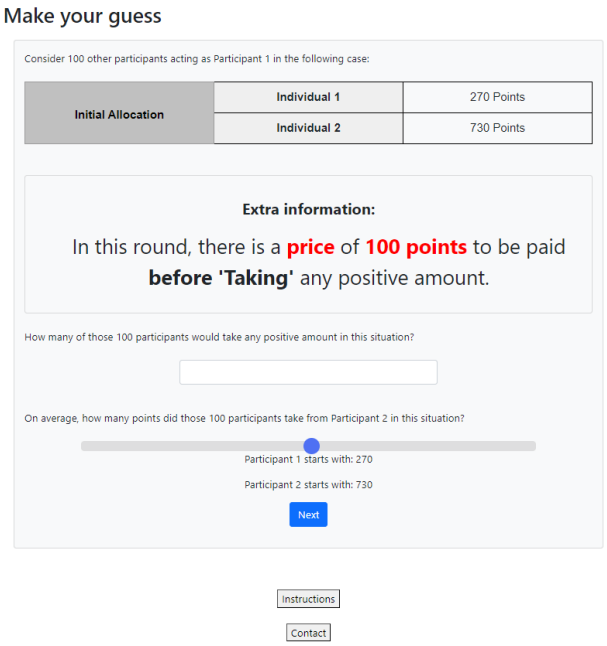

Figure 15: Example: Empirical Expectation

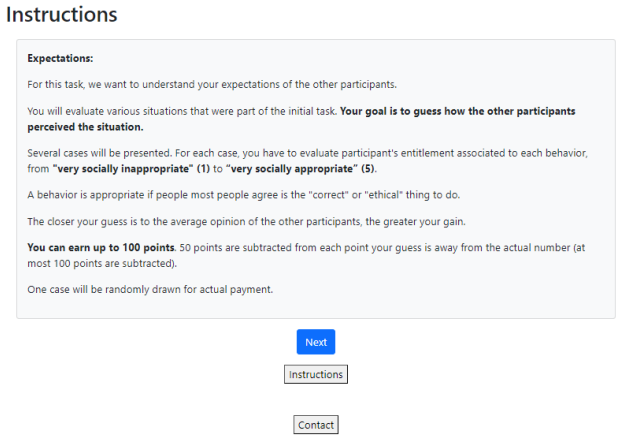

Figure 16: Information - Normative Expectation

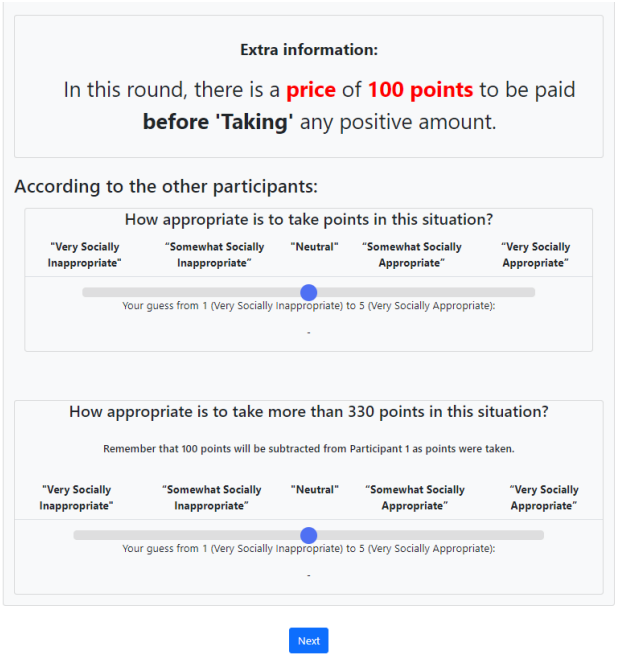

Figure 17: Example: Norm Expectation

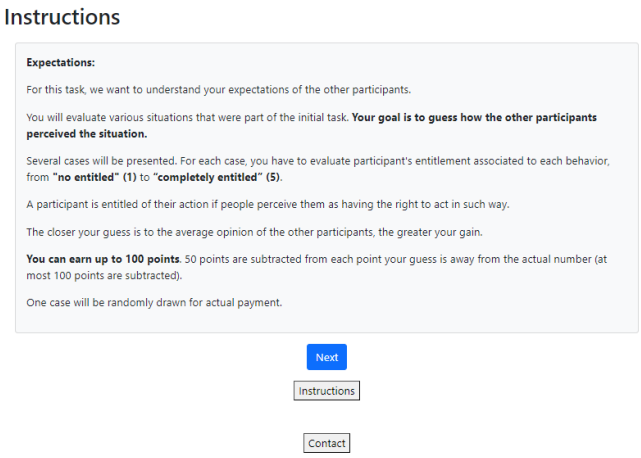

Figure 18: Information - Entitlement

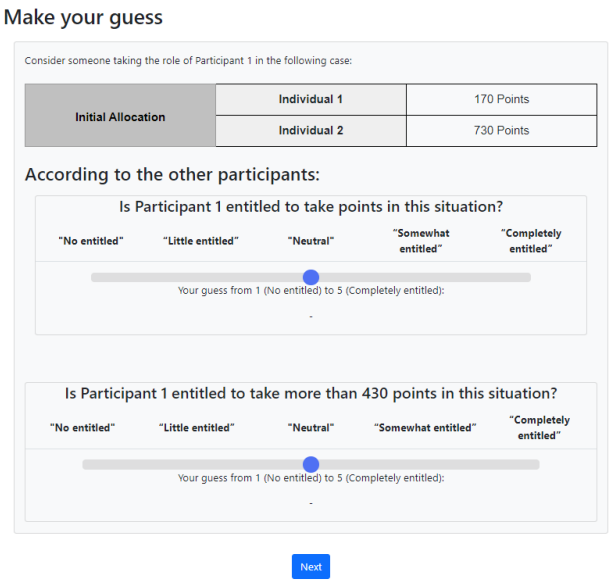

Figure 19: Example: Entitlement