The Impact of Monetary Penalties on Prosocial Motivation: Unveiling Crowding-Out (in) Effects and the Role of Contextual Change

Rafael Teixeira

October 19, 2023

Abstract

Monetary penalties are employed across various contexts and formats to deter undesirable behaviors, yielding complex and often contradictory consequences. 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 money more intensively 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 conclusions. Finally, our study demonstrates that monetary penalties trigger shifts in social norms. These shifts can partially explain the observed crowding-out (in) effects but cannot explain the differences between "fine" and "fee."

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, Simon Gächter and others for their valuable comments and suggestions.

1 Introduction

Monetary penalties are frequently used to enforce laws and rules, aiming to influence human behavior by attaching financial consequences to undesirable actions. The underlying economic logic is to increase the costs of certain behaviors, potentially reducing their prevalence. Surprisingly, in some instances, these penalties can 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 and 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, we do not know the actual behavioral impact of monetary penalties on behavior.

Moreover, penalties are applied in various ways and settings. For instance, environmental legislation frequently incorporates both licenses (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, research conducted by Sunstein (2003); Eriksson, Strimling, Andersson, and Lindholm (2017); Falk and Szech (2013); Bowles (2016) suggests 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 dictator games. Our 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 preferences across settings and gain a clearer understanding of potential crowding-out(in) effects. This contributes to the existing literature that has tested similar phenomena on a penalty setting (e.g., Frey and Oberholzer-Gee (1997); Kurz, Thomas, and Fonseca (2014); Kornhauser, Lu, and Tontrup (2020); Kimbrough and Vostroknutov (2016)) and attempts to reconcile the coexistence of these two seemingly contradictory theories.

Secondly, we aim to test the differences between 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, similar to what was demonstrated by 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 target behaviors.

Finally, we aim to comprehend the underlying mechanism driving behavioral changes. Specifically, we investigate whether the implementation of a monetary penalty triggers distinct social norms, consequently leading to varied behaviors. We use the terminology established by Bicchieri (2016, 2005); 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 elucidated 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 employed to rationalize and elucidate such behaviors. We adapt the methodology proposed by Krupka and Weber (2013) to partially capture this entitlement feeling.

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 engage in multiple rounds where both parties 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 rounds with and without a penalty, allowing us to discern differences in behavior across conditions and identify precise shifts in their choices. In different groups, we implement two subtly different monetary penalties. In one condition, participants face a penalty paid after the "unwanted" behavior, resembling a fine. In the other 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 significance of different settings and 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, we analyze changes at both the extensive margin (cases in which money was taken) and the intensive margin (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. Moreover, 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.

When analyzing social norms as potential mechanisms behind behavioral change, we observe a significant positive relationship between empirical and normative expectations and decisions on both the extensive and intensive margins. For example, on average, a participant who expects fewer people to take any money is less likely to take money than another who expects more people to behave this way. Moreover, the implementation of monetary penalties induces changes in these expectations. Participants 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. 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.

The paper is structured as follows: Section 2 presents the experimental design, Section 3 includes the theoretical analysis and hypotheses, 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 modified the 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 try 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 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 is realized only at the end of the experiment, and the participants do 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:

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

Cases/Twin 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 in which the Dictator starts with fewer points than the Receiver and two in which the Dictator starts with more points. The points are divided into either 900 or 1000 points, and Dictators can take between 800 to 350 points, and on average 570 points. This design is intended to create the possibility of controlling for income effects, 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 a randomized case in each round. To check for a potential order effect, different sessions face different orders, with some sessions starting with the treatment condition and other sessions 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, and there is text reminding them about 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 amount is greater than zero, 100 points are subtracted from the final outcome, otherwise, they keep the 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 to allow 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 version, retaining key elements while attempting to keep 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, thereby 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 conditions of others, 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 specific moral burden associated with those words that could prime individuals and confound the analysis, making it impossible to disentangle what is driven by the word. 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 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 one decoy). 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

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

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 Hypothesis

This section is divided into three parts. The first part describes the theories and hypotheses on crowding-out (in) effects and how they relate to our experiment. The second part discusses the differences between fees and fines. The last part explores social norms as potential mechanisms behind any behavioral changes.

3 .1 Crowding-Out(In) effects

Monetary penalties are commonly used to influence behavior, aiming to reduce the prevalence of undesirable actions. The rational choice theory describes that individuals and businesses weigh the expected costs and benefits of their actions. As monetary penalties increase the cost associated with engaging in undesirable behavior, they potentially lead to a reduction in such behavior (Becker (1968)).

However, the impact of incentives does not always correspond to rational choice theory. Titmuss et al. (1970) theorized that introducing monetary compensation for blood donation might reduce donations. This hypothesis was tested by Mellström and Johannesson (2008), who found mixed results, including a drop in blood donations among female participants when monetary rewards were offered. A similar study conducted by Frey and Oberholzer-Gee (1997), using a survey to analyze support for a nuclear waste storage facility, observed a decrease in support when monetary compensation was introduced compared to the cases with no compensation. Likewise, Gneezy and Rustichini (2000b) showed that offering small monetary rewards led to reduced performance on various tasks, including logical exams. These instances highlight cases where additional rewards resulted in diminished behavior, contrary to what was expected.

Similarly, Gneezy and Rustichini (2000a) described that the implementation of a fine in a daycare for parents who were picking their kids up late led to an increase in the number of parents picking their kids up late. In an empirical study, Earnhart and Friesen (2023) analyzed penalties associated with wastewater discharge and showed that increasing the severity of a penalty might be counterproductive when certainty is low. Such cases exemplify situations where possible monetary penalties actually lead to an increase in the behavior.

These behaviors exemplify the crowding-out theory (e.g., Gneezy and Rustichini (2000a); Frey and Jegen (2001); Frey (2000); Frey and Oberholzer-Gee (1997)), which indicates that the inclusion of a new extrinsic incentive might deteriorate the prosocial concerns, making people act less prosocially. In our setting, such theory suggests that the introduction of a monetary penalty results in an increase in the number of people taking points or the number of points taken.

However, we have to be careful when directly applying such a theory in our context. 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:

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

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

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

Given the penalty, the agent faces three options depending on different β levels: The first option is that the agent does not care about the inequality and keeps taking everything. The second option is that, due to the loss in efficiency ($-p$), it may be more advantageous to maintain the initial inequality, as it results in a larger total amount. The final option is that the agent wants to minimize inequality and redistribute the efficiency loss among the participants, with the agent taking an additional $\frac{p}{2}$

To illustrate the final possibility with numerical examples, consider an agent in the 200/800 scenario. In the control condition, someone with a strong inequality aversion would take 300 points, resulting in an equal distribution of 500/500. Now, with the introduction of the penalty, the agent loses 100 points to take money, leading to a new situation of 100/800.

In this new situation, the agent would take 350 points, resulting in a distribution of 450/450, but taking 50 points more than in the previous case.

That is, some agents would take the money 'more intensively.' This behavioral change could be naively described as a crowding-out effect without any actual change in the prosocial concerns.

However, in our experiments, we use twin cases to control for this income effect/efficiency loss. The 200/800 points case in the treatment condition, after applying the penalty leads to 100/800. Its twin case has exactly the same values, and in the control condition should lead to the same set of decisions, with 450/450 division and 350 points being taken. If the agent increases the amount of money in this condition, it would indicate a clear deterioration of the prosocial situation, as there are other aspects of the decision that are not reflected only by outcomes.

To further distinguish the situations in which a crowding-out effect can be explained by efficiency loss, we define a strict crowding-out of prosocial motivation as:

Definition (Strict Crowding-Out of Prosocial Motivation): Given two initial endowments for the dictator, $x \geq \hat{x}$, and the endowment for the receiver, y, if $\operatorname{argmax} U(x+t) = t^*$ and $\operatorname{argmax} U(\hat{x} + t) = \hat{t^*}$, a strict crowding out of prosocial motivation occurs when $\hat{t^*} > t^*$.

In our experiment, $x \geq \hat{x}$ represents the difference between the treatment condition (x) and the control condition ($\hat{x}$) in the twin cases. For each twin case, in the treatment condition, participants start with an initial endowment that is 100 points higher than their twin in the control condition. However, this initial difference is later subtracted after the 100-point penalty is applied leading to ($x = \hat{x}$).

A strict crowding-out effect refers to an agent taking more money even when the dictator's initial endowments remain the same or greater. This concept is consistent with various prosocial theories. For instance, in Social Value Orientation theory (e.g., Murphy, Ackermann, and Handgraaf (2011)), the angle representing an individual's degree of prosociality decreases from situation x to situation $\hat{x_1}$, indicating a shift towards more selfish behavior. Similarly, in models similar to the inequality aversion model (e.g., Andreoni and Miller (2002)), such behavior would correspond with a decrease in the parameter reflecting inequality aversion[3]. In fact, this definition would directly represent a decrease in the prosocial concerns on any utility function/model that only uses the final outcomes as parameters for the agents' decisions (e.g. Fehr and Schmidt (1999); Bolton and Ockenfels (2000); Andreoni and Miller (2002)).

This concept of 'strictness' allows us to focus on cases where the relationship between output sets and prosociality is more apparent. We acknowledge that our definition of 'strict' crowding-out(in) effects may not encompass all possible instances in which crowding-out may happen. For example, a crowding-out effect might occur if agents increase both their own share and the other's share but proportionally less than expected based on their initial utility function. However, such cases could be due to misspecifications of their utility function's initial parameters.

We meticulously designed cases to create identical sets of possible outcomes, ensuring that money taken in each of the twin cases could be controlled for income effects. This

[3]It is not possible to observe this behavior in a model like Fehr and Schmidt (1999) as such a model only allows for two possible results (50-50 or everything). However, any other model that adapts inequality aversion to a wider range of options would describe this change.

guarantees that a rational and consistent agent would exhibit the same behaviors across various conditions and situations, with potential outcomes being the same. Any utility function solely based on outputs would consistently result in the same amount being taken (t). Therefore, such theories would not anticipate crowding-out effects in this context. Hence, any observed changes would indicate a genuine alteration in prosocial motivation. Furthermore, this setup allows for direct comparisons across different conditions, as the inclusion of fee and fine mechanisms introduces the same fixed cost associated with the action of taking any amount of points, ensuring participants face the same potential outcomes in both cases and again the behaviors, for such theories would predict the same behaviors.

On the other hand, behaviors based on rule-following, as described by Kimbrough and Vostroknutov (2016, 2018), suggest that people tend to follow rules even if doing so goes against their monetary interests. For example, participants are willing to wait at simulated traffic lights showing a red signal even though they would benefit from completing the task more quickly, and there are no penalties for violations. If this is the case, the monetary penalty could be perceived as a new rule that should be followed, leading some participants to drastically reduce the amount taken to conform to the new rule, leading to potential crowding-in effects and an increase in prosocial motivation.

We can also define the crowding-in of prosocial motivation occurs when an agent demonstrates a willingness to allocate more resources to others, even in situations where there are equal or fewer resources available. In such cases, there is an increase in the parameter that represents inequality aversion or social value orientation (SVO):

Definition (Strict Crowding-In of Prosocial Motivation): Given two initial endowments for the dictator, $x \leq \hat{x}$, and the endowment for the receiver, y, if $\operatorname{argmax} U(x+t) = t^*$ and $\operatorname{argmax} U(\hat{x}+t, y-t) = \hat{t^*}$, a strict crowding in of prosocial motivation occurs when $t^* > \hat{t^*}$.

In our experiment, there is another aspect that requires further consideration associated with this potential crowding-in effect: the agents who stop taking money due to the efficiency loss. Models such as Fehr and Schmidt (1999); Bolton and Ockenfels (2000); Andreoni and Miller (2002) offer similar insights about those agents. Individuals who take larger amounts for themselves are less likely to stop taking money. This is because if participants are taking larger amounts of money, they prioritize their self-interest over the others and are less likely to be affected by the penalty[4]. If agents are taking larger amounts of money in the control condition and cease taking money after the introduction of the penalty, these behavioral shifts can be potentially attributed to crowding-in effects.

Notice that crowding-in and crowding-out theories describe opposite behaviors, with one reflecting an increase in prosocial motivation and the other indicating a decrease. Both theories cannot coexist simultaneously 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. Confronting these two theories is crucial for a better understanding of the impact of monetary penalties.

Our setup can directly pinpoint any crowding-out effect, providing a clear test of how individuals might become less prosocial in such a context. We can also analyze potential crowding-in effects and how people might strictly adhere to rules even when they consis-

[4]In Appendix A, we provide an example of a quadratic function for inequality aversion to illustrate one possible threshold calculation.

tently would not. By potentially examining both crowding-out and crowding-in effects and pinpointing the exact behavioral changes, we can engage in this confrontation among the theories.

As this section outlines, multiple theories propose various potential outcomes for monetary penalties. To simplify our hypotheses and minimize writing a list of potential alternative explanations, our primary hypotheses are based on what traditional economic theory and outcome-based models (e.g. Fehr and Schmidt (1999); Andreoni and Miller (2002)) would predict when comparing the twin cases. This approach enables clear predictions as we control for income effects. However, for a more comprehensive understanding of all behavioral changes, our data analysis includes descriptions and indications of behavioral changes both with and without controlling for income effects.

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

As mentioned earlier, in both the treatment and control conditions, the agents would take the same amount of points because they encounter an identical set of potential outcomes. However, with the introduction of a monetary penalty, a fixed cost is added, which may discourage some agents from taking any points due to the efficiency loss associated with paying the penalty. As a result, the total amount taken is expected to decrease. Based on these expected changes, we can extend this hypothesis to cover both the extensive margin (whether to take any points) and the intensive margin (how many points to take, if any).

For the extensive margin:

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

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

Hypothesis 2a suggests that some individuals would stop taking any points as the penalty introduces an efficiency loss. Hypothesis 3a examines the behavior of people who are 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 are taking 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 who are taking larger amounts, for example, all 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 as potential crowding-in effects.

For the intensive margin:

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

Hypothesis 4a describes that, given the twin cases, the set of potential outcomes is the same and behavior should be the same.

In general, crowding-out effects would predict a deterioration of the behaviors, leading all hypotheses to describe an increase in the number of times and amount of money taken.

To observe crowding-in effects, we also analyze individual changes and highlight cases in which drastic amounts of money were taken in the control condition (hypothesis 3a), but the participant ceased to take in the treatment condition.

3 .2 Fine vs. Fee

Monetary penalties and financial obligations are used in diverse ways in different cases. For example, to prevent many people from parking in specific places, locations establish a parking permit fee, but in other places, they issue parking tickets as fines. Environmental agencies may impose an annual fee and grant licenses to industrial facilities to cover the cost of pollution. On another hand, when companies violate environmental regulations, they may face fines as a financial penalty.

These incentives are used to prevent or decrease the prevalence of specific behaviors. In economic terms, they all aim to increase the relative cost of such behavior, making it less beneficial and thereby decreasing its prevalence. Even though they have the same economic purpose and would be characterized similarly in economic theory, the different formats might lead to different perceptions, changing the perceived moral implications and leading to other psychological interpretations of the same behavior, as described by Sunstein (2003) and Bowles (2016). Similarly, Falk and Szech (2013) describe that the market setup deteriorates the prosociality of its participants, causing people to act more selfishly.

Our goal is to analyze how small changes in the implementation of monetary penalties can lead to different behavioral outcomes, even though they may represent economically equal scenarios. This initial understanding of the impact of monetary penalties on behavior paves the way for more effective and targeted interventions, improved legal frameworks, and enhanced public policies. It represents a crucial step toward shaping a society that not only punishes wrongdoing but also promotes cooperation and prosocial behavior, and analyzing which is the most efficient to perform penalties.

To achieve this goal, we created two stylized versions: a fee (a monetary penalty paid before taking money) and a fine (a monetary penalty paid after taking money). Our aim was to make them as economically similar as possible while simplifying some aspects of both the fine and the fee, but highlighting crucial differences.

Firstly, our fine does not include a risk component or the feeling of getting caught. To implement a risk component and ensure comparable expected values between the fee and fine, the cases would have to undergo significant changes across the conditions. This would make it impossible to create comparable cases across conditions while controlling for income effects simultaneously.

Secondly, to simulate any fee condition, we must create a scenario in which the monetary penalty is paid before the action, establishing a two-stage decision. Implementing a fee without this initial decision is not possible. In contrast, in the typical implementation of fines and control conditions (situations with no penalties) in a more realistic scenario, this first-stage decision does not exist. In our opinion, this difference represents a fundamental distinction across conditions. Future research will attempt to disentangle this aspect from the payment by itself. However, our current goal is to examine potential differences between versions of fees and fines, capturing key components of their distinctions while controlling for crucial aspects that would not allow us to precisely compare cases or theories. We highlight this difference as a reason for variations across conditions and how these cases would differ in a real-life scenario.

Meanwhile, it is worth noting that both fees and fines represent the same cost-efficient

change in our conditions. Both impose a fixed 100-point cost aimed at creating a barrier to take money. In economic terms, both fees and fines represent the same set of potential outcomes, and, given that, the change is primarily a framing effect. The implementation only creates a change in the perception of commitment and timing of payments. Given this, traditional economic theory or even models similar to inequality aversion (Andreoni and Miller (2002)) that only consider the set of potential outcomes would predict that fees and fines would not differ.

However, these different perceptions could lead to distinct behavioral changes. For example, various payment methods can elicit different responses, even when the amounts involved are the same (Zellermayer (1996)). By making the payment more salient in the first stage of the fee condition, people might be less likely to take money. Another possibility is that the timing of the penalty might make the moral aspect of the decision more salient (Eriksson et al. (2017); Ellingsen and Mohlin (2022)), further reducing the perceived appropriateness of taking any money, which could decrease the number of people choosing to take money. If these conditions hold true, it suggests that fees could result in more significant crowding-in effects compared to fines.

On the other hand, Gneezy and Rustichini (2000a) and Gneezy et al. (2011) discuss the concept of entitlement that individuals may experience after paying for an action, which can contribute to the crowding-out effect. The timing of payment and the first-stage decision could increase the feeling of entitlement for the fee in comparison to the fine. Similarly, participants might perceive taking money as more socially acceptable because they have "paid for it" through the fee. If these conditions hold true, it would suggest that fees could result in more significant crowding-out effects compared to fines.

Hence, different theories could imply that the fee would lead to larger crowding-out effects but also stronger tendencies to follow rules. To maintain consistency across our hypotheses, we base our main hypothesis on classic economic theory, which predicts no differences across fee and fine. In the previous section, all hypotheses were illustrated with a number and the letter 'a.' All the previous hypotheses will be re-analyzed to determine if there are significant differences between fees and fines. Consequently, all hypotheses also have a 'b' version that describes:

Hypotheses #b: There are no differences across fee and fine.

3 .3 Social Norms and Perceived Entitlement

As previously described, a monetary penalty is intended to increase the cost associated with a specific action and decrease its prevalence. However, different and unique behaviors (e.g., Gneezy and Rustichini (2000a); Gneezy et al. (2011)) have been observed in which the incentive backfires, leading to a deterioration of the situation known as crowding-out effects. Several theories have attempted to explain how and why such behaviors might occur, and economic theory (e.g., Bénabou and Tirole (2006, 2003); Janssen and Mendys-Kamphorst (2004); Frey and Oberholzer-Gee (1997)) generally highlights two crucial aspects: incomplete information and strategic interactions.

One possible scenario, for instance, is when a new incentive is used as a coordination tool in situations with multiple equilibria. For example, people might perceive the penalty as new information that many others are engaging in the specific behavior, and as people do not know which is the "correct equilibrium," the penalty leads them to coordinate their actions towards the same behavior (Janssen and Mendys-Kamphorst (2004)). Another possibility is

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

The dictator game exhibits two interesting characteristics, which we leverage in our study: Firstly, it provides a direct and clear measure of prosocial preference. Models such as those proposed by Fehr and Schmidt (1999), Andreoni and Miller (2002), Yang, Onderstal, and Schram (2016), and others frequently use this setup to exemplify prosocial concerns directly.

Secondly, the dictator game minimizes the roles of strategic interaction and incomplete information, often invoked by these theories to describe potential crowding-out effects. This is because participants have complete information about the outcomes, and there is no direct interaction with other participants. Moreover, the experiment is entirely anonymous, preventing participants, other 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. Then, our design places emphasis on the role of other potential explanations[5].

On the other hand, psychological and new economic theories (e.g., (Frey & Jegen, 2001; Capraro & Perc, 2021)) explain how incentives can influence individuals' engagement with a situation and might lead to changes in other aspects of decision-making. For example, people might trigger different moral concerns (e.g., (Capraro & Perc, 2021; Ellingsen & Mohlin, 2022)) when the penalty is implemented, or different situations might evoke distinct self-image concerns (e.g., (Tonin & Vlassopoulos, 2013)).

Similarly, regarding rule-following tendencies and potential crowding-in effects, (Kimbrough & Vostroknutov, 2016) emphasizes that people are willing to forgo personal gains to follow rules and highlights the role of norms in shaping prosocial preferences. People tend to conform to norms and rules, and a fine introduces a new rule where such behavior is not desired, prompting individuals to follow it even when it is not individually beneficial.

(Kimbrough & Vostroknutov, 2016) argues that social preferences are primarily influenced by social norms, aligning with similar arguments made by (Andreoni & Bernheim, 2009) and (Krupka & Weber, 2013). There is an extensive body of literature describing how social norms can affect behavior (e.g., (Bicchieri & Dimant, 2019), (Bicchieri, 2005), (Bicchieri, 2016), (Cialdini & Trost, 1998)), and models such as (Krupka & Weber, 2013) and (Akerlof & Kranton, 2000), along with experiments like (Xiao & Bicchieri, 2010), demonstrate that people tend to conform to social norms, and norms can be directly incorporated into preferences.

Building on this literature, we believe that introducing a new incentive might activate different social norms. As individuals tend to conform to these new norms, their behavior may undergo potential changes. Social norms have been defined and described in various ways, but we adhere to the description associated with (Bicchieri, 2005), which distinguishes norms based on 'empirical expectations' (beliefs about what others do) and 'normative expectations' (beliefs about what others think we ought to do).

In this context, our goal is to capture conformity to these norms, implying a positive monotonic relationship between norms and behavior. If this holds true, an increase/decrease in perceived norms should result in an increase/decrease in the corresponding behavior.

[5]Hence, if crowding-in or crowding-out effects are observed even in this simplified scenario, it provides strong evidence that they could be even more pronounced in other, more complex scenarios.

In addition to social norms, entitlement is considered a potential factor in explaining crowding-out effects (e.g., (Bénabou & Tirole, 2006; Gneezy et al., 2011)) as people might use this to motivate their choices. Therefore, it would be valuable to explore this potential channel. While it is not feasible to directly measure this sense of entitlement through incentivized means, we have created a new measure to partially capture this motivation. We adapted the methodology developed by (Krupka & Weber, 2013), utilizing a coordination game to incentivize the question of what the participant believes the group's opinion is.

Our methodology also draws inspiration from the field of social psychology, particularly attribution theory (e.g., (Peterson et al., 1982; Dykema, Bergbower, Doctora, & Peterson, 1996)), which examines how individuals perceive the causes and motivations behind everyday experiences, constructing possible explanations through social inferences based on the context and individuals involved.

While the social psychology literature lacks incentivized methods for measuring such motivational aspects, we believe that by adapting the methodology proposed by Krupka and Weber (2013), we can partially capture the social construction of motivation (in this context - entitlement). Future research can delve deeper into these methods to capture aspects associated with concerns related to social image and motivated reasoning, as discussed in previous studies (e.g., Epley and Gilovich (2016)).

We also would like to test if the fine and fee lead to different social norms, and hence lead to different types of behavior. This relation across norms and framing effects is vastly presented in the literature, but not always directly explored. (Ellingsen, Johannesson, Mollerstrom, & Munkhammar, 2012) describe how labeling the prisoners' dilemma differently leads the participants to expect higher or lower levels of coordination for the other participants, and adjust their behavior in response. (Krupka & Weber, 2013) describes that different descriptions of a dictator game (regular vs. take-give) can lead to different normative expectations, which can lead to different behaviors. Similarly, our hypothesis is that the fee and the fine would trigger different behaviors, for example, the fee might lead to higher levels of perceived entitlement, which could trigger different behaviors.

To further illustrate this discussion, consider a general utility function that incorporates all types of social norms. The agent's utility, U, depends on their gains x (initial endowment), t (amount taken), and p (penalty), as well as the social norms associated with the amount taken, denoted as $N(t, t_{\text{emp}}, t_{\text{nor}}, t_{\text{ent}}|\text{framing})$. Such norms are directly integrated into their utility function, with t_{emp} representing empirical expectations, t_{nor} normative expectations, and t_{ent} perceived entitlement. Notice that the norms are conditional on the framing, that is, fee and fine might lead to different norms. We also include a parameter γ, representing the agent's propensity to conform to the norms:

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

We expect a positive relation between the amount taken t and social norms:

$$\frac{\delta U}{\delta t}\frac{\delta t}{\delta t_{\text{emp}}} > 0, \quad \frac{\delta U}{\delta t}\frac{\delta t}{\delta t_{\text{nor}}} > 0, \quad \frac{\delta U}{\delta t}\frac{\delta t}{\delta t_{\text{ent}}} > 0$$

Hence, the higher the empirical expectations, normative expectations, and entitlement, the higher the amount taken. If the monetary penalties trigger higher/lower norms, the amount taken would be higher/lower. If the fee and fine trigger higher/lower norms, more/-less money is taken on the fee condition compared to the fine. These discussions are sum-

marized by the following hypotheses:

Hypothesis 5a- Changes in behavior are positively associated with social norms/entitlement.

Hypothesis 5b- Changes between the fee and the fine are positively associated with social norms/entitlement.

4 Results

The study involved 201 participants, 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 (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 observe no significant difference across the order as observed in Appendix 10, and hence all the corresponding treatment sessions are grouped together for data analysis. Appendix 9 shows 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 extensive and intensive margins, while analyzing the behavioral differences between the fine and fee. 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:

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

We aim to explain the amount taken ($Take$) by individual i in round r. β_0 captures the mean behavior of the control section in the fine condition. The variable *Fine* is a dummy for the fine treatment condition, 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 condition, 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 effect model to control for individual differences, and the residuals are clustered at the individual level.

[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.

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:

	(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 observation, (2) twin cases

The results are very similar with or without controlling for the income effect. Using regression (2) as our main benchmark, we observe a statistically significant decrease in the amount taken in the fee condition (-27), supporting hypothesis 1. In contrast, there is a slight non-significant decrease in the fine condition (-6). When comparing the impact of the fee and fine treatments, we find a marginally significant difference ($\chi^2(1) = 2.89, p = 0.0894$).

Result 1a: *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.

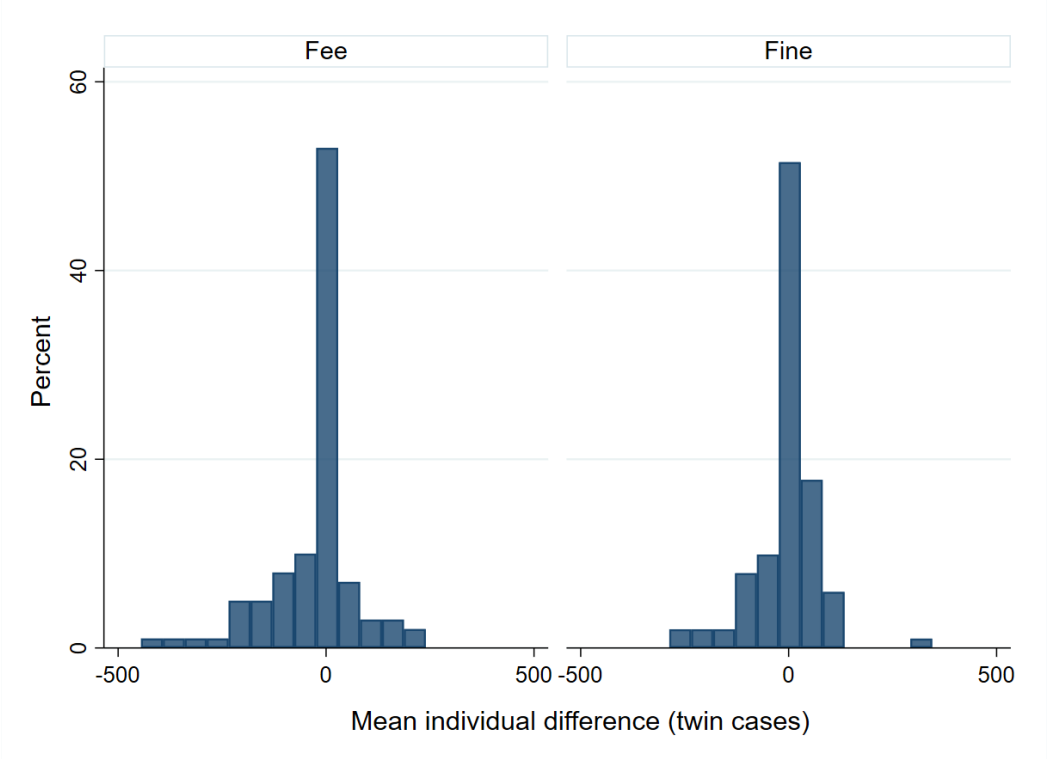

Figure 1: Distribution of mean individual changes for each condition

The behavior of the majority of participants remains relatively consistent, taking the same amount of money in both treatment and control conditions, aligning with predictions from models similar to Andreoni and Miller (2002). However, we also observe significant deviations among some individuals. One participant, for instance, 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 analyze two contrasting effects: some individuals cease taking points altogether, while others exhibit notable changes in the intensity of their takings. This opens the discussion for the subsequent hypotheses related to the intensive and extensive margins. We commence with an analysis of the extensive margin, as outlined in our hypothesis 2, which posits that monetary penalties should lead to a decrease 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 that money is taken - (3) all observation, (4) twin cases

The observations provide evidence supporting hypothesis 2, as 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 test for differences between the fee and fine treatments and analyze the 10 percentage points difference in impacts using a chi-square test ($\chi^2(1) = 5.01, p = 0.0252$). The results indicate significant differences between the fee and fine treatments.

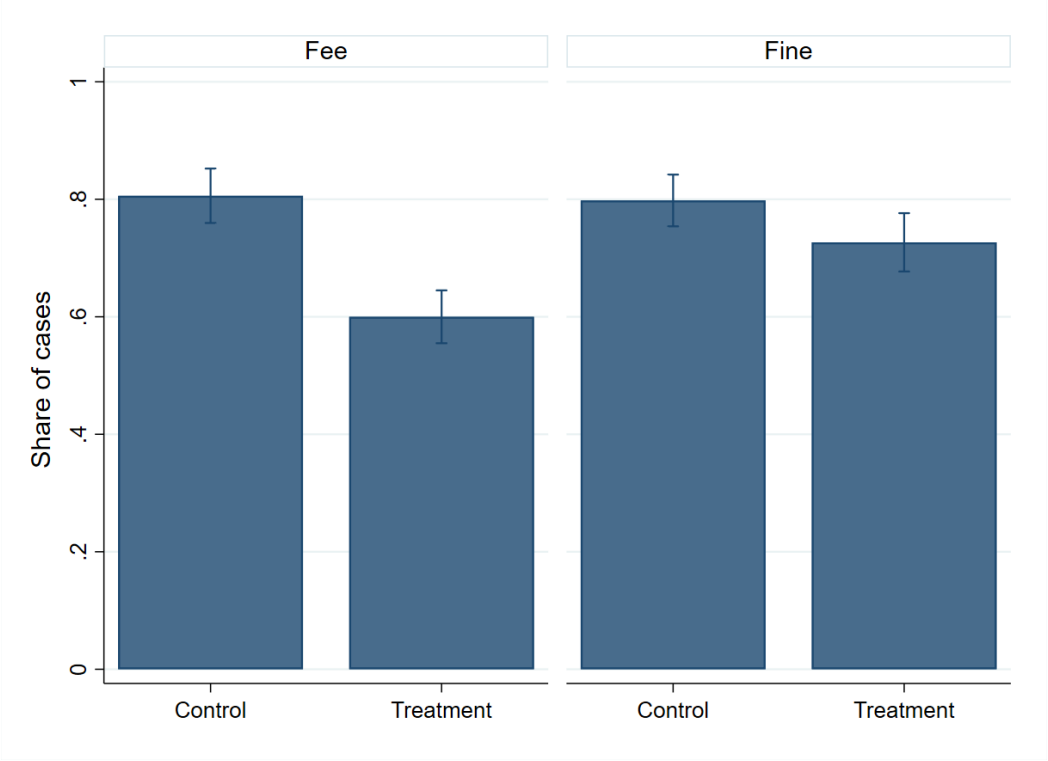

Figure 2: Predicted value of the share of cases in which money is taken and their 95% confidence interval in each condition (Twin Cases).

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

Result 2a: *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: *There are significant differences 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 what those agents did 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 ($take + endowment/total$) that the dictator earned in the control condition for the situations in which no money was taken in the treatment condition, Figure 4 shows the distribution of amount taken for the same respective cases.

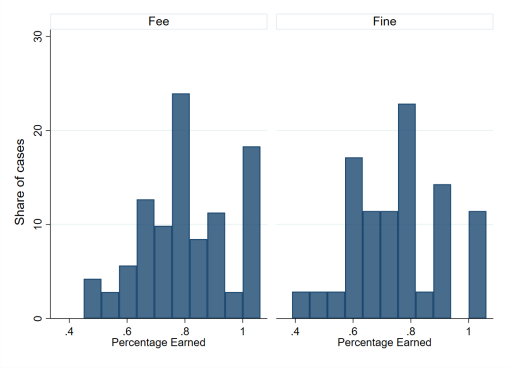

Figure 3: Distribution of the share that the dictator earned in the control condition by those who did not take money in the treatment condition.

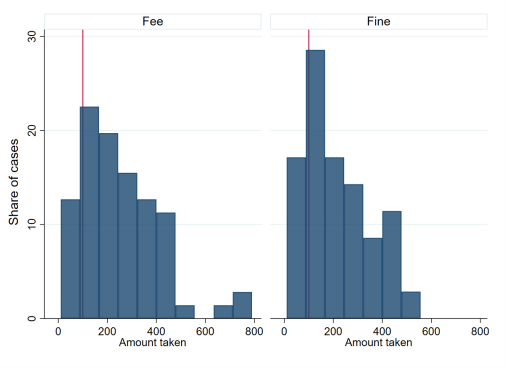

Figure 4: Distribution of the amount taken by the dictator among those who did not take money in the treatment condition.

On average, dictators obtain around 80% and 77% of the total available for those specific cases under the fee and fine conditions, respectively. In some instances, the dictator obtains larger amounts or even the entire available amount, but then ceases to take any money after the implementation of the penalty. Specifically, in the fee condition, the dictator obtain 100% of the money in 18.3% of cases, while in the fine condition, this occur in 11.43% of cases, and these individuals decided to stop take any money after the implementation of the penalty.

Additionally, we compare the amount taken to directly compare it with the monetary penalty (100-points). Among those cases, the fee leads to an average reduction of 248 points, and 200 points for the fine condition, with no differences among the treatment conditions ($\chi^2(1) = 0.88, p = 0.3482$). As a benchmark criterion, we compare the amount taken with 100 points associated with the cost of the monetary penalty, and the average amount taken is shown to be significantly different ($\chi^2(1) = 42.50, p = 0.0000$). To illustrate further, in approximately 50% of the cases, the 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 are evidence of a crowding-in effect.

Result 3a: *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 proceed with our analysis of the intensive margin.

Before we start our analysis, it is important to clarify the sample used in each regression. Generally speaking, the intensive margin looks at the participants who took any money, as represented in regression (5). However, it is expected that the participants who took any money in the treatment and control conditions would be different, potentially leading to an

endogenous effect.

However, we can control this aspect by pinpointing the individual changes. That is, we examine the amount taken by participants who have taken money in the treatment condition and compare it with their respective control case. Hence, keeping the same participants and the same cases for both conditions. Regression (6) presents the results for the entire dataset. Regression (7) focuses on the twin cases, controlling for the individuals and income effects.

Here, it is important to highlight the significance of the coefficient, *ControlDiff*, since the fee and fine conditions might select different participants who are willing to take money. If this is the case, *ControlDiff* will capture the mean difference across these individuals. Table 5 describes the details:

	(5)	(6)	(7)
	Take	Take	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. After controlling for income effects, 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.

Regression (6) also reveals 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 4a: *Conditional on taking money in the treatment condition, the implementation of both the* ***fee*** *and the* ***fine*** *increased the amount taken.*

Result 4b: *No significant differences between* ***fee*** *and* ***fine*** *increase on the amount taken.*

In summary, our findings highlight the significant and heterogeneous impacts of introduc-

ing 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. Interestingly, 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 observed some differences in the impacts of different cases and the relationship between inequality and behavioral changes are discussed and illustrated in Appendix D.

4 .2 Social Norms and Entitlement

In this section, we investigate three potential mechanisms behind the behavioral changes: empirical expectations, normative expectations, and perceived entitlement. We hypothesize a positive monotonic relationship between the analyzed behaviors and norms/entitlement, meaning that if something is perceived as more expected/appropriate/entitled, individuals are more likely to behave accordingly. To achieve this, we first analyze whether the introduction of a monetary penalty affects the measures of social norms and entitlement. Subsequently, we check whether the observed behavioral changes can be attributed to potential changes in such measures by controlling for the treatment effects with respect to social norms/entitlement.

For each measure of social norms/entitlement, we elicited two different 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.

The second aspect relates to the intensive margin. For empirical expectations, we initially asked about the average amount of money that the same 100 people would take. Notice that this question describes the aggregate amount taken, not the intensive margin by itself. To create a better proxy for the intensive margin, we weighted the expected amount taken by the expected number of participants who would take money in each case, thereby creating the amount of money being taken conditional on money being taken. For normative and entitlement aspects, we asked participants to indicate the appropriateness/perceived entitlement for taking approximately 70% of the total amount.

The regressions are illustrated in Table 7, with regression (11) describing a linear regression with random effects for empirical expectations, regressions (12) describing the weighted empirical expectations, and regressions (13)-(14) describing the regressions for normative expectations and entitlement.

For these regressions, as we expect to analyze the impact on the intensive margin and capture the effect of the crowding-out effect, we examine the norm change for those agents who continue to take money in the treatment condition. That is, we analyze the norm change for the sample used in the intensive margin of the previous session. In the Weighted Empirical Expectations, in a few cases, participants argued that zero participants would take money, and we cannot create its weighted version:

	(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 lead to changes in the general empirical expectations of the amount taken; however, the weighted empirical expectations show a significant increase for the fine condition and a large increase for the fee condition, though not significant due to the high variance associated with this new measure. Additionally, both the fee and fine conditions lead to (marginally) significant increases in the perceived appropriateness levels of taking larger amounts of money. Interestingly, the fee condition leads to an increase in the perceived entitlement to take larger amounts of money, while it has no impact on the fine condition.

To conclude our analysis, we incorporate social norms and entitlements into similar regression models as in the previous sections to investigate whether changes in social norms/entitlements 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 a new regression to examine the new treatment effects after adding the social norms/entitlement. Specifically, we first run the same regression as before for the four observations (twin cases 2 and 4):

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

Subsequently, we conduct the following regression:

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

Empi, *Norm*, *Enti* represent the empirical expectation, normative expectation, and entitlement, respectively. As we check for intensive and extensive margins, the extensive margin uses *Participation* as the dependent variable.

If β_3, β_4, and β_5 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 take larger amounts. After the regression, we test 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 regression[7]. In regression (16), we add the measures for social norms/entitlement to understand their contribution to behavioral changes. Regressions (17) and (18) reproduce the same results for the intensive margin (Take) using linear regression.

	(15) Participation	(16) Participation	(17) Take	(18) 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		0.00488***		0.684***
		(0.000662)		(0.0439)
Normative		0.00712***		1.564**
		(0.00195)		(0.718)
Entitlement		0.00339**		1.516**
		(0.00170)		(0.672)
Constant	0.815***	0.139**	395.4***	47.91**
	(0.0242)	(0.0539)	(15.94)	(20.80)
N	804	804	556	556

Standard errors clustered at the individual level in parentheses

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

Table 8: Channels: replicate extensive margin (15) and norms controls (16), replicate intensive margin (17) and norms controls (18)

First, regressions (15) and (17) almost perfectly replicate the results of regressions (4)

[7]To make the comparison more consistent.

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.

Secondly, the coefficients for the social norms and entitlement are positive and significant for all conditions. This indicates that measured social norms can partially explain the 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 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 5a: 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 5b: 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).

5 Discussion

We examine the influence of monetary penalties on prosocial motivation within the context of a dictator game, contrasting stylized versions of a fine vs. a fee. Our aim is to explain theoretical contradictions that describe both a crowding-out effect (e.g., Gneezy and Rustichini (2000a)) - which leads to a deterioration of prosocial motivation - and a crowding-in effect (e.g., Kimbrough and Vostroknutov (2016)) - potentially resulting in an increase in prosocial motivation. Ultimately, this research aims to generate a nuanced understanding of human behavior and provide insights for policy implications.

Even though monetary penalties are widely employed in various settings to deter undesirable outcomes, their actual impact is not always clear. There are several examples illustrating crowding-out effects (e.g., Gneezy and Rustichini (2000a); Frey and Jegen (2001); Gneezy et al. (2011)), suggesting that penalties might backfire by deteriorating individuals' prosocial concerns. On the other hand, Kimbrough and Vostroknutov (2016, 2018), and others describe that people have a tendency to follow rules, and in the presence of a penalty, individuals are likely to adhere to it even at the significant expense of their own gains. Such observations lead to contradictory predictions and findings that cannot coexist, raising questions about the true impacts of penalties.

We are able to observe both crowding-in and crowding-out effects and provide explanations for seemingly contradictory observations: The monetary penalty has heterogeneous impacts on different individuals.

Firstly, a significant portion of the participants exhibit consistency, and models similar to outcome-based models (e.g., Andreoni and Miller (2002)) regarding inequality aversion or other measures of prosocial motivation would be consistent with these observations. We create a situation in which the agents face the same set of possible outcomes with and without the penalty, controlling for income effects. Such models would predict that the agents would act very similarly with and without the penalties, and we observe a large share of participants taking the same amount across conditions.

Secondly, when the penalty is introduced, many participants refrain from taking money, even when they had previously taken larger amounts. It was expected that some participants would refrain from taking money, as the penalty induces an efficient loss that might affect some participants. We also observe that the participants were consistently taking substantial amounts, which can hardly be explained only by the efficiency loss. The participants consistently took values above 200 points, and in some cases, they took all the available money. This drastic reduction in the amount taken by the participants is evidence of a rule-following tendency, similar to what was described by Kimbrough and Vostroknutov (2016), and can be interpreted as a crowding-in effect.

Thirdly, conditional on continuing to take money after the implementation of the penalty, it is observed that the participants significantly increase the amount they take. Hence, the participants see the penalty as a motivation to act more selfishly and take money more intensively. This behavior exemplifies a deterioration of prosocial motivation and provides clear evidence of a crowding-out effect.

It's worth noting that we do not observe crowding-out effects in the sense described by Gneezy and Rustichini (2000a), which refers to a change at the extensive margin. However, their setting includes various other factors, such as a more significant impact of incomplete

information and strategic interaction, where agents can coordinate towards other equilibria (e.g., Janssen and Mendys-Kamphorst (2004)). Furthermore, their fine introduces an income effect, which can partially explain their results, as discussed in the theory section. It is possible that in specific settings, a crowding-out effect might occur even at the extensive margin; however, we were not able to observe this version of crowding-out effects.

Meanwhile, our observations of crowding-out effects are highly robust, as they consistently manifest within a conservative setting. In the twin cases, which we use to control for efficiency loss/income effects associated with the penalty, the crowding-out effects appear as nothing more than a shift in framing. Participants encounter the same set of possible outcomes in both scenarios. Our use of 'crowding-out' adheres to the strict crowding-out effect in our definition, representing a clear example of a change in prosocial concerns. This holds true when considering models analyzing output, such as those presented by Andreoni and Miller (2002) or Murphy et al. (2011).

Moreover, the dictator game, which minimizes strategic interaction and incomplete information—often the primary driving factors behind crowding-out effects (e.g., Bénabou and Tirole (2006) and Frey and Jegen (2001)) also reduces the influence of potential other factors that could lead to behavioral changes. This highlights that crowding-out (or in) effects can be observed even in simpler settings.

The crowding-in and crowding-out effects create opposing forces that influence the impact of the monetary penalty. Different settings and situations might cause one effect to prevail over the other, leading to conflicting observations. This is partially exemplified by the differences between fees and fines.

We implemented fees and fines to make them as natural and directly comparable as possible. The fee imposes the penalty before the action while also creating a first-stage decision where participants must choose whether to take money or not. The fine is subtracted after taking any money.

We observe that, at the aggregate level, the fine is inefficient and does not significantly change the average amount taken in the experiment. This does not necessarily indicate that the fine does not impact behavior; rather, it suggests that the crowding-in and crowding-out effects balance each other, resulting in a null impact at the aggregate level. In fact, participants who continued taking money increased their amount taken by almost 4%, which offset the 5% point decrease in the number of cases in which money ceased to be taken due to the implementation of the fine.

In contrast, the fee condition is able to make an impact, leading to a decrease in the average amount of money being taken, resulting in a reduction of almost 9%. This different aggregate result is reflected in the almost 15% point drop in the number of cases in which money is taken, while the participants who continue taking money increased the amount taken by around 6%.

The differences between the aggregate results for the fee and fine are primarily driven by the extensive margin. The fee results in significantly fewer people taking money compared to the fine. Meanwhile, the changes at the intensive margin are not significantly different. The behavior in the control conditions for both the fee and fine is very similar. This further reduction in the number of participants taking money can be attributed to a greater tendency to follow the rules promoted by the fee condition. Therefore, people with potentially fewer social concerns are respecting the fee, which can also be associated with a greater crowding-in effect.

Hence, the results indicate the heterogeneous impacts of monetary penalties, with some participants being consistent, others following rules and leading to crowding-in effects, while

others use this as an opportunity to take more money, leading to crowding-out effects. The balance of these forces depends on the context, and the fee condition proved to be more efficient than the fine, as the fee was able to reduce the aggregate amount taken, while the fine was inefficient.

The results also prompt avenues for future research. Firstly, a deeper exploration of the relationship between the severity of penalties and their influence on crowding-in and crowding-out forces. It is likely that higher penalty levels reduce the frequency of occurrences, but it is also possible that they amplify crowding-out effects.

Further investigations could also delve into the identification of individuals more likely to exhibit crowding-out or crowding-in behavior in response to monetary incentives. By identifying those more inclined to act in specific ways, penalties can be better tailored to their target.

Additionally, the potential for a more nuanced comprehension of interventions and policy behavioral designs, as proposed by Bowles (2016), deserves exploration. For instance, environmental legislation often employs fines as a means to deter environmental damage, alongside the establishment of licenses (similar to a fee) to permit specific behaviors. Furthermore, contemporary approaches, such as the use of 'carbon markets,' may seem analogous in economic theory, but they can elicit different behavioral responses. For example, carbon markets might align with Falk and Szech (2013)'s findings, suggesting that the implementation of a market can erode moral and prosocial behavior. A thorough comparison and contrast of such policy tools are essential for designing more effective policies and interventions.

We also endeavor to explore potential mechanisms behind the observed behavioral changes. Specifically, we investigate the role of social norms and perceived entitlement. Social norms are recognized as a key component in explaining behaviors (e.g., Krupka and Weber (2013); Bicchieri (2005, 2016)), and together with entitlement, they are often associated with crowding-in or crowding-out behavior (e.g., ? (?); Gneezy and Rustichini (2000a); Janssen and Mendys-Kamphorst (2004); Gneezy et al. (2011); Kimbrough and Vostroknutov (2016)).

To capture the concept of entitlement, we developed a novel measure by adapting the methodology originally proposed by Krupka and Weber (2013). Our aim was to create a measure that could partially capture this motivational aspect. Drawing inspiration from Krupka and Weber (2013), we utilized a coordination game designed to incentivize participants to consider what they believe the group's opinion to be.

This approach is rooted in social psychology, specifically in attribution theory (e.g., (Peterson et al., 1982; Dykema et al., 1996)), which explores how individuals perceive the causes and motivations behind everyday experiences. Attribution theory tries to understand how people construct explanations through social inferences based on the context and individuals involved, and how those perceived motivations shade behavior. While the field of social psychology lacks incentivized methods for precisely measuring motivational aspects like entitlement, we believe that adapting the methodology proposed by Krupka and Weber (2013) provides a valuable means to partially capture the social construction of motivation, particularly in the context of entitlement.

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

For the social norms, adopt the terminology developed by Bicchieri (2016, 2005); 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), and we use an adapted version of Krupka and Weber (2013) to measured normative expectations.

We first observe that the implementation of monetary penalties does indeed shape social norms. For example, participants believe that others are less likely to take money when the penalty is in place and consider taking larger amounts of money as more socially appropriate. Interestingly, in the fee condition, there is a higher perceived entitlement when agents take larger amounts, whereas there is no significant change in the fine condition.

This result suggests that people comprehend that taking money in such a situation is morally questionable[8]. However, they also perceive the penalty as a potential motivation to take more money – adopting a mindset of 'if I'm doing something wrong, let me take more benefit', thinking it is more socially acceptable, for example, to take large amounts of money.

Next, we examine whether these behavioral changes can explain shifts in behavior. We observe that norms and entitlement are positively correlated with behavior, both at the extensive and intensive margins. This means that individuals who believe people would take more money or consider taking larger amounts more socially appropriate are also more likely to take more money themselves. This result indicates conformity to social norms; people act in accordance with their beliefs.

Lastly, we observe that changes in social norms can partially account for these behavioral shifts, partly moderating the coefficients for the treatment effects. Consequently, we find evidence that alterations in norms can partially elucidate the crowding-in and crowding-out effects.

These results provide direct evidence that crowding-out and crowding-in effects can be partially attributed to social norms, as theorized and described by various authors such as Gneezy and Rustichini (2000a); Gneezy et al. (2011), Kimbrough and Vostroknutov (2016), and Ellingsen and Mohlin (2022). This finding is particularly significant in the context of a dictator game, which illustrates that social concerns permeate individual decision-making. In the dictator game, agents do not engage in direct strategic interaction with the experimenter or other participants. While social concerns have been previously observed in the dictator game, as shown in Andreoni and Bernheim (2009), our results reveal that even small manipulations can lead to significant changes in prosocial motivation. This partial explanation sheds light on these previously controversial behaviors, highlighting the change in social norms as an individual concern (as in Krupka and Weber (2013) or Akerlof and Kranton (2000)) that leads to such behavior.

These findings carry substantial implications for policy interventions. They align with discussions in works such as Bicchieri and Dimant (2019), Sunstein (2003), and Lane, Nosenzo, and Sonderegger (2023). These studies delve into how nudges, interventions, and laws can influence social norms, thus impacting behavior. However, our study underscores the need for careful execution of such interventions, as they can also lead to unintended crowding-out effects.

Furthermore, it is crucial to comprehend the intricate interplay between context and norms, as exemplified in our experiment. The implementation of both fees and fines in our setting results in the same set of potential outcomes, highlighting a framing effect. However, the context has proven sufficient to trigger different norms, and these norms can partially

[8]This also indicates that the experiment design is indeed interpreted as a rule, a fine, or a monetary penalty.

account for the observed behavioral changes. Social norms are notoriously resistant to change (e.g., Dimant, Van Kleef, and Shalvi (2020); Bicchieri and Dimant (2019)), making it challenging to directly influence them. Yet, the relationship between norms and framing opens up opportunities for indirect interventions.

Understanding this dynamic allows us to indirectly target norms by strategically altering the context. By thoughtfully designing environments and contexts, policymakers can exert influence over the development and transformation of social norms, thereby fostering desired behavioral changes. Changing the context can trigger different norms, ultimately leading to changes in behavior. This nuanced approach offers a promising avenue for effectively achieving policy objectives while acknowledging the inherent resilience of social norms.

However, social norms alone cannot comprehensively explain the behavioral differences between the fee and fine conditions. These distinctions between the conditions may be influenced by other factors embedded in our experimental design. The most notable divergence between the fee and fine conditions emerges at the extensive margin, suggesting that the first-stage decision may function as a commitment device, eliciting distinct feelings and emotions.

It's worth noting that this commitment device does not substantially alter the underlying social norms, so the social norms associated with the fee and fine conditions may not differ significantly. Therefore, another intriguing possibility is related to Zellermayer (1996)'s discussion of the 'pain of paying.' The first-stage decision renders the payment more salient and triggers diverse emotional responses that may not be adequately captured by social norms.

One possibility could be driven by calculated greediness, as discussed in Rand, Greene, and Nowak (2012). In the fee condition, participants spend more time contemplating their decision, as observed in appendix A, which might lead to changes in the decision-making process without necessarily triggering alterations in social norms.

It's also worth noting that we observe small differences in the impact of the fee and fine considering the initial inequality, as described in appendix D. The results indicate that the fee is more effective in reducing the number of people taking money, especially when the agent is ahead. This suggests that there is an interaction between the situation and the format, and these differences are triggered in specific conditions. Future research might also aim to further analyze other aspects of the context to better understand the behavioral impacts.

One possibility is that the agent only starts to care about this decision when they have moral reasons for doing so, which are not necessarily salient when the agent is behind. The interaction between the two-stage decisions and this initial doubt might trigger this behavioral change.

Future research could delve even deeper into unraveling the various motivations and rationales underpinning this behavioral change. Understanding the interplay between commitment mechanisms, emotions, and social norms can provide valuable insights for designing interventions and policies that target behavioral shifts more effectively.

6 Conclusion

Monetary penalties are commonly used to discourage undesirable behavior, but their precise impact is often unclear as observed by the existence of conflicting theories regarding potential crowding-out and crowding-in effects. Our results explain why there are such contradictory

theories by showing that monetary penalties have heterogeneous impacts on individuals. Many individuals use the penalty as a motivation to cease the undesirable behavior, even when they were engaging in it intensively, thus exemplifying crowding-in effects. Meanwhile, others start to engage in the behavior more intensively, exemplifying a crowding-out effect.

The balance of these forces deeply depends on the context and situation, and we illustrate this by highlighting the differences between a fee and a fine. In our study, the aggregate results for the fine show these forces balancing out, illustrating a case in which the penalty might seem inefficient. However, when the same penalty is introduced before the action, as a fee, it further decreases in cases in which the target behavior is observed, proving to be efficient. Thus, we demonstrate how small contextual changes can affect the efficiency of penalties.

Lastly, we observe that the introduction of monetary penalties shifts perceived social norms. People expect fewer others to engage in the target behavior when a penalty is in place, while they also believe that engaging in the behavior intensively is more socially appropriate when the penalty is implemented compared to when there is no penalty. These changes in social norms can partially explain behavioral changes, accounting for both crowding-in and crowding-out effects. Hence, we provide direct evidence that social norms can partially explain crowding-out and crowding-in effects. However, social norms do not address the differences between a fee and a fine.

References

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

Andreoni, J., & Bernheim, B. D. (2009). Social image and the 50–50 norm: A theoretical and experimental analysis of audience effects. *Econometrica*, *77*(5), 1607–1636.

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.

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.

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.

Dimant, E., Van Kleef, G. A., & Shalvi, S. (2020). Requiem for a nudge: Framing effects in nudging honesty. *Journal of Economic Behavior Organization*, *172*, 247–266.

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.

Earnhart, D., & Friesen, L. (2023). Certainty of punishment versus severity of punishment: enforcement of environmental protection laws. *Land Economics*, *99*(2), 245–264.

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. (2023). 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.

Murphy, R. O., Ackermann, K. A., & Handgraaf, M. J. (2011). Measuring social value orientation. *Judgment and Decision Making*, *6*(8), 771–781.

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.

Rand, D. G., Greene, J. D., & Nowak, M. A. (2012). Spontaneous giving and calculated greed. *Nature*, *489*(7416), 427–430.

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, U, represents that the agent cares about their initial endowment, x, the amount they take, t, and negatively weights, $\beta > 0$, the inequality between their gains and others' gains, $(x + t) - (y - t)$, on a quadratic scale.

In the control condition:

$$\max : U(t) = x + t - \beta(x + t - (y - t))^2 = x + t - \beta(x - y + 2t)^2$$

In the treatment condition, there is an extra 100 points penalty if the agent takes points. This holds for both fee and fine:

$$\max : U(t) = \begin{cases} x - \beta(x - y)^2 \\ x + t - 100 - \beta(x + t - 100 - (y - t))^2 = x + t - 100 - \beta(x - y - 100 + 2t)^2 \end{cases}$$

By maximizing the control condition, we obtain that the maximum argument is $t = \frac{1+2(x-y)\beta}{8\beta}$, and the maximum is $\frac{1}{16\beta} + \frac{x+y}{2}$. Similarly, in the treatment condition, the agent would keep $\frac{1}{16\beta} + \frac{x+y-100}{2}$ if they decide to take points. However, some agents would stop taking points if:

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

Hence, some agents would continue to take points and keep a proportionally larger total share, while some agents would stop taking points, and the overall outcome would depend on the distribution of β across the population. By solving this inequality for all possible cases, the maximum amount that the dictator would take is 138 points.

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

Table 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 it is apply the fee at that specific session, *Order* is a dummy variable that is equal to 1 if the session started with the treatment condition, for last, there is an interaction term that evaluates that the order effect my 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) illustrate the order effects on the control conditions, using the observations only associated with the control. Regression (Treatments) illustrate 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:

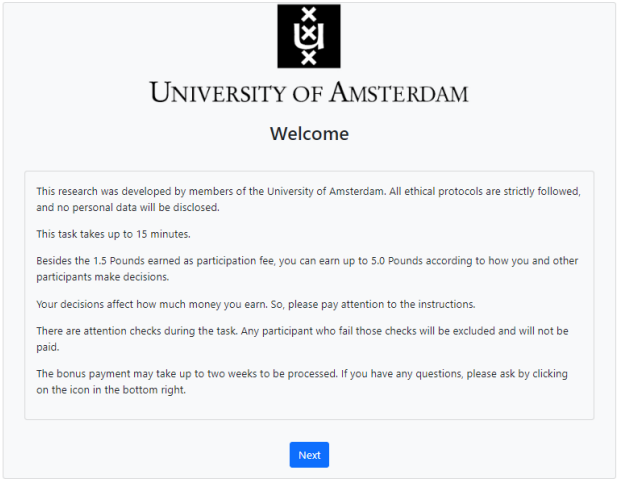

Figure 5: Introduction

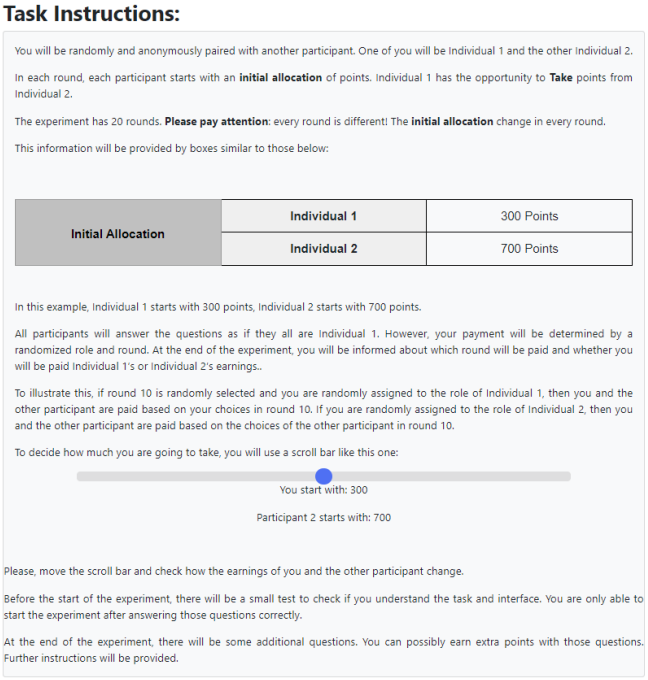

Figure 6: Instructions

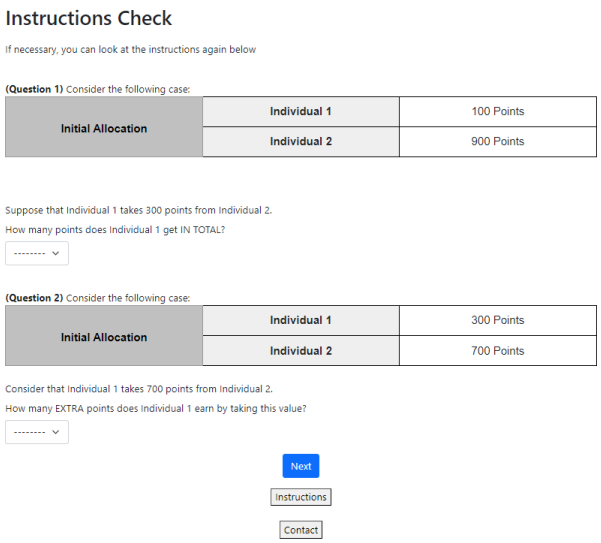

Figure 7: Example - Comprehension check

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

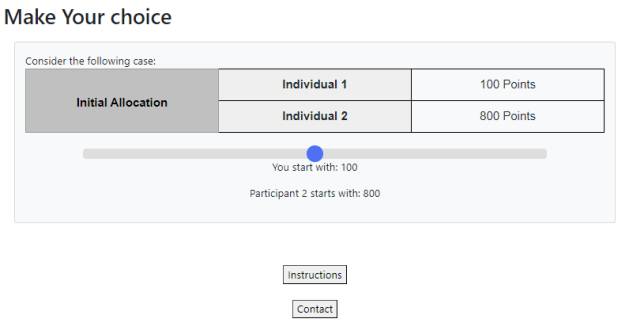

Figure 8: Example: Control Condition

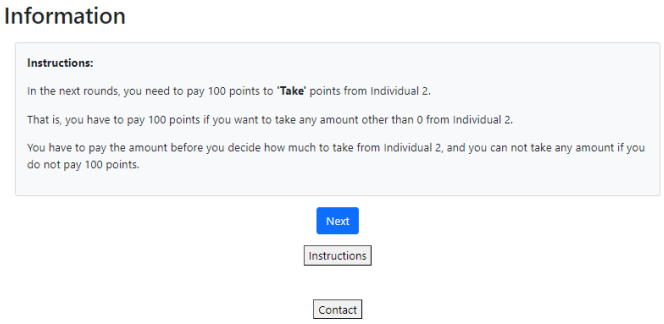

Figure 9: Information - Fine

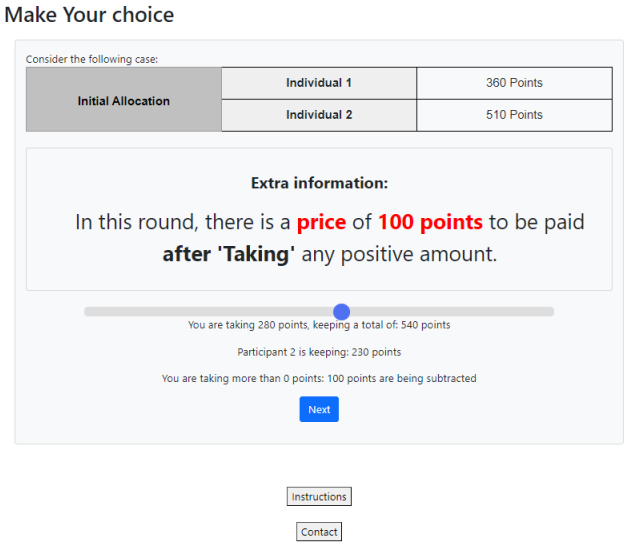

Figure 10: Example: Fine Condition

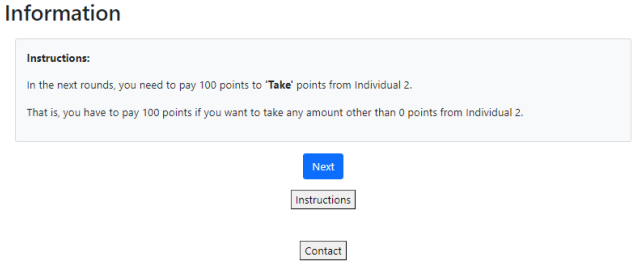

Figure 11: Information - Fee

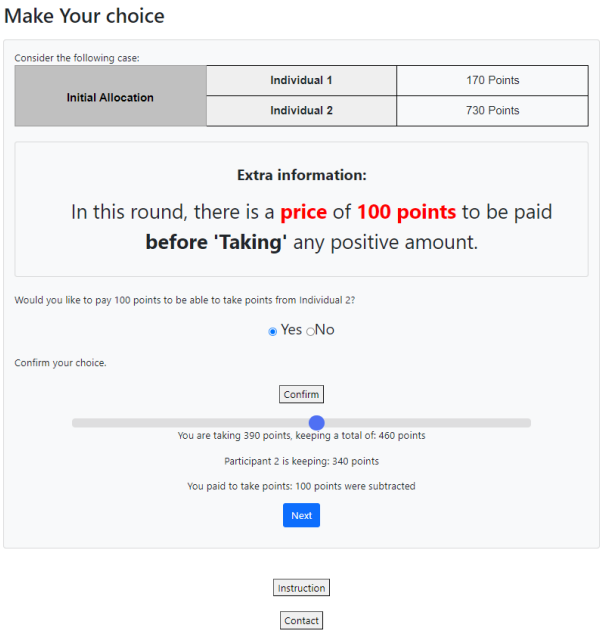

Figure 12: Example: Fee Condition

Social Norms and Entitlement:

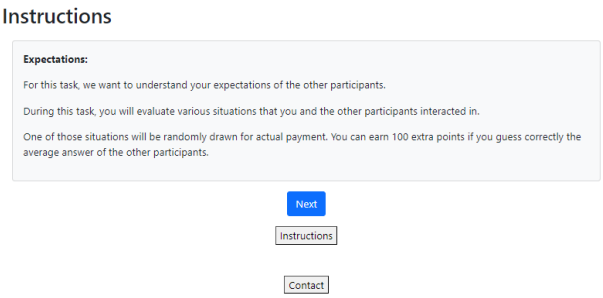

Figure 13: Information - Empirical Expectation

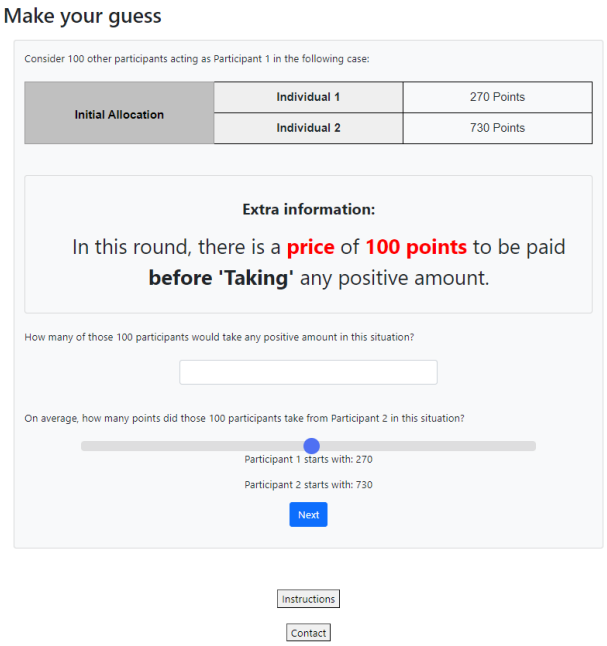

Figure 14: Example: Empirical Expectation

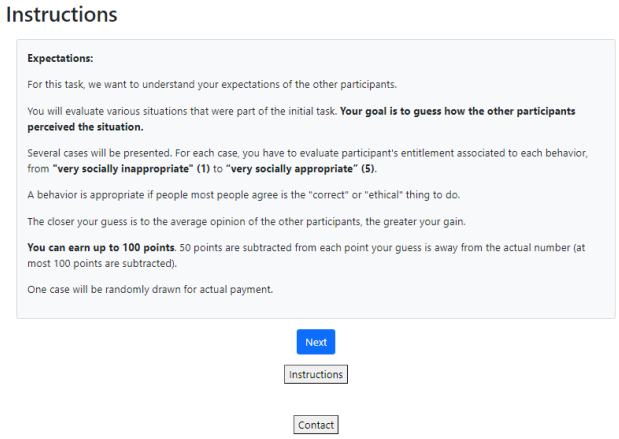

Figure 15: Information - Normative Expectation

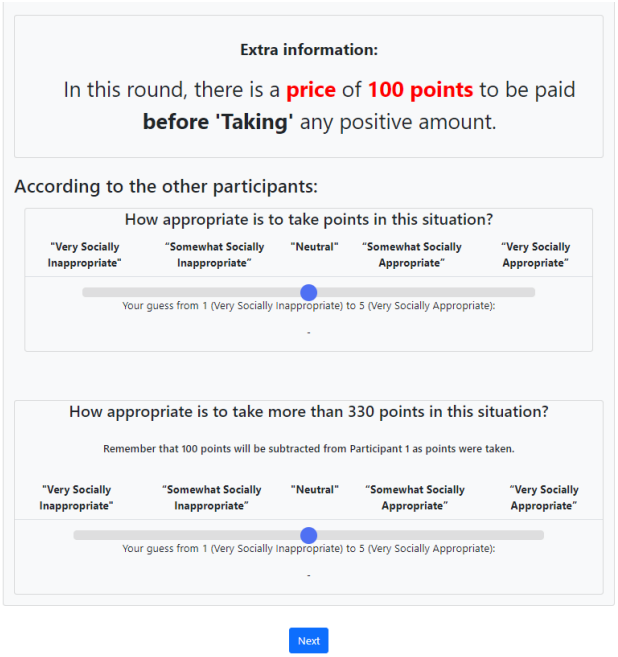

Figure 16: Example: Norm Expectation

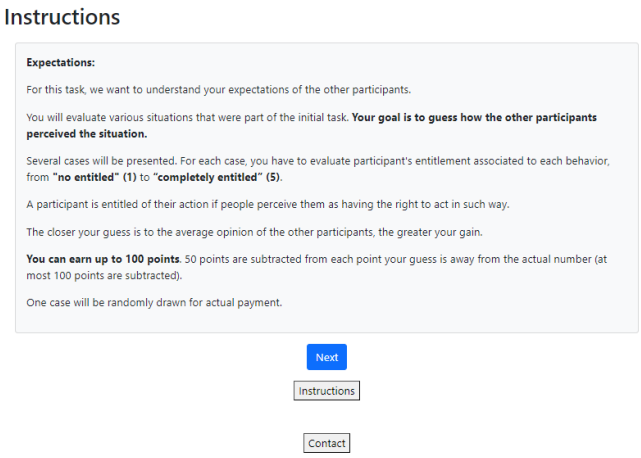

Figure 17: Information - Entitlement

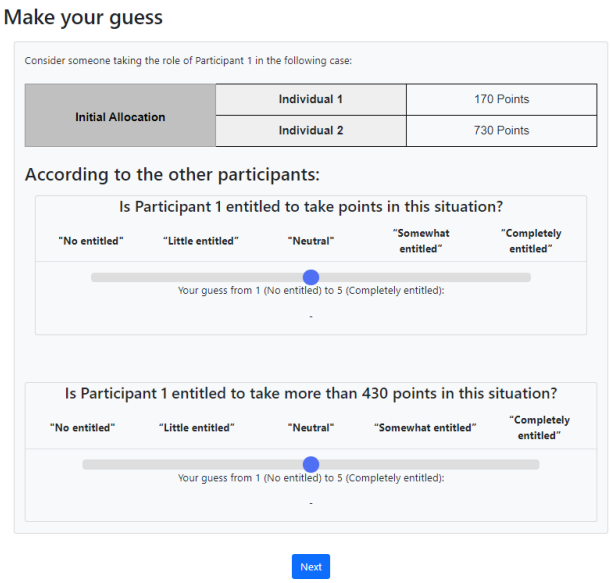

Figure 18: Example: Entitlement