Fines vs. Fees: The Impact of Monetary Penalties on Prosocial Motivation

Rafael Teixeira

November 30, 2023

JOB MARKET PAPER
Most recent version

Abstract

We investigate the impacts of distinct monetary penalties using a modified dictator game that allows participants to take money from others. We introduce a penalty of equal monetary value in two distinct formats: a 'fine,' imposed after taking money, and a 'fee,' paid before taking money. Our findings reveal that the fee is more effective than the fine. In comparison to a situation with no penalty, the fee significantly reduces the aggregate amount taken, while the fine exhibits no significant overall impact. Additionally, we observe heterogeneous effects of the penalties on individuals' prosocial behavior. Some individuals take more money when facing a penalty, indicating a crowding-out effect, while others abstain from taking money when confronted with the penalty, even when they take substantial amounts without penalties, evidence of a crowding-in effect. Overall, the fee proves to be more effective in promoting crowding-in than the fine, while crowding-out effects are similar across conditions. Furthermore, we demonstrate that the implementation of monetary penalties induces changes in perceived social norms. As individuals conform to these norms, such changes partially explain the crowding-out and crowding-in effects.

Keywords: Crowding-out effect, crowding-in effects, 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, Alexis Belianin, Alejandro Hirmas, Shaul Shalvi, Ro'i Zultan, George Lowenstein, Silvia Sonderegger, Chris Starmer, Sander Onderstal, Jan-Willem Stoelhorst, Alexander Vostroknutov, Simon Gächter, and others for their valuable comments and suggestions. I would also like to express my gratitude to the participants of the seminars at the University of Amsterdam, the University of Nottingham, and the University of Paderborn.

1 Introduction

In this paper, we explore the behavioral features of fines and fees, commonly deployed monetary penalties used to discourage "undesirable" behavior. Fees and fines exemplify some of the nuances inherent in the diversity of how monetary penalties are implemented in everyday situations. For instance, environmental legislation frequently employs a combination of fees and fines to deter environmentally harmful behaviors. Governments often issue emission permits or impose fees on companies, granting the privilege to emit a predetermined amount of greenhouse gases. Conversely, companies that violate environmental regulations are frequently subjected to fines as a punitive measure. Traditional economic theory posits that penalties influence trade-offs by increasing the relative cost of "unwanted" behavior and diminishing its prevalence. However, it fails to distinguish between penalty formats, addressing only traditional features such as risk or timing preferences.

We eliminate other confounding factors such as risk concerns, to focus on one essential distinction across fees and fines: Fines are paid *after* the targeted behavior, while fees are paid *before* the targeted behavior. In our experimental setup, both the fee and the fine offer the same monetary incentive. Consequently, traditional economic theory predicts that the behaviors observed across the two conditions should be identical (Tversky and Kahneman (1988)). However, potential distinctions between fees and fines might arise from the fact that penalties can have impacts beyond cost-benefit analysis, as their implementation might lead to changes in prosocial concerns (e.g., Gneezy and Rustichini (2000a); Kimbrough and Vostroknutov (2016); Frey and Jegen (2001)). Our objective is to disentangle the different impacts associated with implementing a monetary penalty, controlling its effects on trade-offs, and consequently exploring its influence on prosocial concerns. This approach allows for a direct test of different theories and empirical observations associated with such potential changes in prosocial concerns (e.g., Bénabou and Tirole (2006, 2003); Frey and Oberholzer-Gee (1997); Kimbrough and Vostroknutov (2016)). Moreover, by understanding potential behavioral differences across fees and fines, we shed light on potential policy interventions, showing that the format of a penalty might affect the penalties' efficiency and generate insights for better legislation.

We also explore the mechanisms influencing such potential changes in prosocial concerns by analyzing social norms. Lane, Nosenzo, and Sonderegger (2023a); Kimbrough and Vostroknutov (2016) demonstrate that implementing a law or rule might induce shifts in social norms. Additionally, a substantial body of literature (e.g., Bicchieri (2005, 2016); Xiao and Bicchieri (2010); Krupka and Weber (2013)) indicates that individuals tend to conform to social norms. Following this rationale, we investigate whether implementing

(different) monetary penalties triggers (different) shifts in social norms. As individuals conform to these norms, it results in behavioral changes. In doing so, we provide additional context for models proposed by Ellingsen and Mohlin (2022) and Bénabou and Tirole (2006, 2003) that describe changes in prosocial concerns, while linking their approaches to models like Krupka and Weber (2013) and Kimbrough and Vostroknutov (2016), which emphasize the importance of conformity to social norms.

Monetary penalties are generally not applied in multiple formats in daily situations, and these different formats are often associated with many confounding factors that also affect the behavior. Hence, aiming to understand the impact of the format itself, fee vs. fine, and to comprehend the influence of social norms, we employ an online experiment. We analyze the decisions made by participants in a modified dictator game. Participants go through multiple rounds in which they start with different initial endowments and have the option to take money from other participants. They make decisions under two different conditions: a control condition in which no penalty is implemented and one of the treatment conditions in which a monetary penalty is introduced.

Taking money is the "bad behavior" that we aim to deter with a penalty. In different groups, we implement one of the following monetary penalties: The 'fine' condition, in which participants face a penalty paid *after* any money is taken, and the 'fee' condition, in which participants face a penalty paid *before* taking any money (i.e., participants have to pay before being able to take money). Both the fee and fine reflect the same monetary value, with the only difference between the treatment conditions being the timing that the penalty is imposed — *before* and *after*.

We analyze three behaviors: at the aggregate level, which involves examining the average amount of money taken by all subjects; at the extensive margin, which refers to the number of instances in which money is taken; and at the intensive margin, which refers to the amount of money taken, conditional on taking money.

Following the choices in the dictator game, we assess the participants' social norms for different situations, aiming to measure norms reflecting the intensive and extensive margins. To capture social norms, we adopt the methodology developed by Bicchieri (2005) and Krupka and Weber (2013), which categorizes social norms into empirical (what others do) and normative (what others should do) expectations. Additionally, we seek to measure "perceived entitlement." Entitlement is a potential explanation for crowding-out effects (e.g. Gneezy and Rustichini (2000a)), as people may feel entitled to engage in the target behavior after "paying for it." This feeling may not necessarily be captured through social norms, as norms might reflect a myriad of different aspects. To quantify

this, we develop a novel methodology inspired by Krupka and Weber (2013), creating an incentivized method to assess entitlement.

In certain instances, implementing a monetary penalty might result in a potential deterioration of the situation and social concerns (Frey and Oberholzer-Gee (1997); Frey and Jegen (2001); Gneezy and Rustichini (2000a)). For instance, a study conducted at a daycare center (Gneezy and Rustichini (2000a)) demonstrated that introducing a fine for parents picking up their kids late unexpectedly led to even more tardiness. This phenomenon has been characterized as a crowding-out effect. In our setting, this would manifest as an increase in the instances or intensity of money being taken when a penalty is implemented compared to the situation with no penalty.

One interpretation of Gneezy and Rustichini (2000a)'s result is that the penalty may make parents feel entitled to be late because they paid for it when they paid the fine. If this is the case, implementing a penalty might increase the appropriateness and perceived entitlement associated with taking money, which would be indicated by our measures in social norms. Moreover, based on this reasoning, it is plausible that a fee paid before the action could exacerbate this sense of entitlement, resulting in an even more pronounced crowding-out effect than a fine.

On the other hand, monetary penalties can lead to a substantial improvement in the situation and prosocial concerns, as people tend to follow rules even when it is costly (e.g., Kimbrough and Vostroknutov (2016, 2018)), leading to potential crowding-in effects. A monetary penalty can be seen as a signal that a behavior is "bad" or a rule indicates that such behavior is "unwanted". In our setting, this would imply that people would decrease the amount of money taken or stop taking money when they were taking substantial amounts of money without the penalty.

Similarly, if the penalty is perceived in such a way, people might believe that fewer individuals will take money, or that taking money is less socially acceptable. Moreover, the fee, which moves the decision on taking money or not before choosing the amount taken, might emphasize such aspects and lead to more crowding-in effects than the fine.

Furthermore, an increase in prosocial concerns, as in crowding-in effects, and a decrease in prosocial concerns, as in crowding-out effects, represent opposing forces and cannot occur simultaneously. By disentangling these effects, our study contributes to the literature that analyzes the prosocial impacts of incentives (e.g., Bénabou and Tirole (2006, 2003); Frey and Oberholzer-Gee (1997); Kimbrough and Vostroknutov (2016)), directly testing both crowding-in and crowding-out effects. Through this, we aim to resolve the paradox that penalties have been found to lead to both crowding-out and crowding-in

effects.

The findings reveal systematic differences between the fee condition and the fine condition and illustrate the heterogeneous impacts of monetary penalties on behavior: some participants show crowding-in effects, an increase in prosociality, while others display crowding-out effects, a decrease in prosociality.

At the aggregate level, the fine leads to no significant impact on the amount taken compared to the control, suggesting that this penalty was not effective. In contrast, the fee results in a significant reduction in the aggregate amount taken compared to the control.

At the intensive margin, participants consistently take more money in both the fine and fee conditions compared to the same decisions in the control condition. This increase in the amount taken suggests a crowding-out effect, with participants becoming less socially concerned after the implementation of the penalties. We observe no significant differences in the crowding-out effects between the fine and fee conditions.

In contrast, at the extensive margin, both the fine and fee conditions result in a reduction in the number of instances where money is taken compared to their respective control conditions. Moreover, the fee condition leads to a significantly greater reduction than the fine. Consequently, the fee promotes more prosocial behaviors than the fine, indicating a stronger crowding-in effect.

We also observed that many participants take larger amounts of money in the control conditions but stop taking any money when penalties are implemented. In some cases, participants take all available money in the control condition but completely stop taking any money once a monetary penalty is introduced. We observe no significant difference between the fee and fine in terms of the average number of points that people stop taking. However, these substantial reductions provide further evidence of crowding-in effects.

Hence, we observe that the fee is more effective than the fine, and this difference reflects the heterogeneous indirect impact that the different penalty formats have on behavior. Some people 'use' the penalty to act less prosocially, providing evidence of crowding-out effects, while others 'use' the penalty to act more prosocially, offering evidence of crowding-in effects. The crowding-out effects are consistent across conditions, whereas the fee leads to higher levels of crowding-in than the fine.

When analyzing social norms as potential mechanisms behind behavioral change, we observe that the implementation of monetary penalties induces changes in expectations. Participants, for example, compared to situations with no penalty, believe that fewer individuals would be willing to take money with the implementation of penalties, but they also perceive taking large amounts of money as more socially appropriate when a

penalty is in place. Intuitively, the logic seems to be: "You should not do it, but if you do, you should make the most of it."

We also find that social norms can partially account for the treatment effects at both the intensive and extensive margins, thereby partially explaining the crowding-out and crowding-in behaviors. However, we find no evidence that changes in social norms explain the differences between fees and fines.

The paper is structured as follows: Section 2 presents the experimental design, and Section 3 is 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]

Participants interact in a dictator game in which a participant (the Dictator) decides how much money to take from another participant (the Receiver). We modified the standard dictator game into this taking game to capture the impact of implementing a monetary penalty on an 'undesirable behavior.' The original dictator game incentivizes giving behavior, which is generally viewed positively. By reframing the game in terms of taking, we aimed to model a situation where such behavior is likely 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 was realized only at the end of the experiment, and the participants did not directly interact at any time.

These 20 dictator games are divided into two blocks under two different conditions. In one block with 10 different cases reflecting various initial endowments, the participant makes decisions in a control condition. In this condition, the participant is informed that

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

they can take points from the other participant without any further information. In the other block containing 10 decisions with the same cases, the participant makes decisions in one of two treatment conditions. In these treatment conditions, the participant is informed that there is a 100-point penalty associated with taking any money. Therefore, the impact of each monetary penalty is observed within-subjects, while the differences across the different monetary penalties are observed between-subjects.

We vary the order of the control and treatment decisions across experimental sessions, with some sessions starting with the control condition and others starting with the treatment conditions. We do this to investigate if the order of the treatment might affect behavior. The order of the different cases is randomly presented to the participant. The 10 cases and their different initial endowments are described in Table 1.

Twins	**Cases**	**Dictator's Endowment**	**Receiver's Endowment**
1	1	100	800
	2	200	800
2	3	170	730
	4	270	730
Decoy 1	5	360	510
3	6	500	400
	7	600	400
4	8	550	350
	9	650	350
Decoy 2	10	630	310

Table 1: Cases - the cases represent the 10 different initial endowments for the dictator and receiver in various rounds of the dictator game. Twins reflect a difference in endowment for the dictator equal to the size of the monetary penalties (100 points), and they are used to control for income effects associated with the penalty. Decoys represent cases without twins but with a different total amount being divided.

The cases encompass a range of diverse initial endowments, including scenarios where the dictator begins with more money than the receivers and instances where the dictator starts with less money than the receiver. In some cases, the dictator starts with a higher endowment than the receiver, while in others, the receiver starts with more endowment. We do this variety to check the robustness of any behavioral change across initial inequality.

The endowments aim to generate twins and enhance decision robustness. Participants consistently allocate either 900 or 1000 points, fostering decision consistency and contributing to behavioral change robustness. In twin cases, the Dictator starts with

an extra 100 points in the second case of each pair, enabling control for income effects arising from the monetary penalty introduction. Illustrating this, take 'Twins 1' as an example: In case 2, under the treatment condition, the participant begins with a 200/800 initial endowment. Upon taking any money and encountering the associated monetary penalty, there is a reduction of -100 points in their endowment, resulting in a scenario with 100/800. This aligns with the same numbers observed in case 1, its twin.

Therefore, by comparing the decisions across the twin cases as described above, we can control for the income effect associated with the penalty. This principle applies to any twin. By controlling for the income effect associated with the penalty, we can account for changes in the trade-off, thereby focusing on the potential impacts on prosocial motivation

The decoy cases are included to provide participants with some variety, preventing them from facing decisions with the same value repeatedly. For such cases, we cannot control for income effects.

In all decisions, 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. This setup remains consistent in both the control and treatment conditions.

We have two treatment conditions, implementing the 100 points penalty in two different ways:

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

In the fee condition, the deduction of 100 points occurs *before* the participant makes their decision. The participant is presented with the following question: "Would you like to pay 100 points to be able to take points from Individual 2?" If the participant chooses to pay the fee, 100 points are subtracted from their endowment, and the slider is activated, allowing them to decide on the allocation.

Before the start of the blocks with the treatment decisions, the participant is informed that for the next decisions, there is a penalty associated with 'taking' any money. In each decision screen, in addition to the information described above, participants in the treatment conditions are reminded about the penalties. The specific text for each treatment condition can be found in Table 2:

Treatment Condition	Text informed to the participants
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.
Control	No additional text

Table 2: Treatments - text provided to the participant in the decision screen for each treatment condition, Control, Fee & Fine. The only difference is related to the timing of the decision.

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

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

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

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.

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, twins 4, and decoy 1). For each possible

mechanism, one case was randomly selected for payment. Participants could earn an additional 100 points if their answers matched the group average. To maintain consistency and avoid confusion across the measures, we employed a linear rule to determine points earned based on the distance from the correct answer for all measures.

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

To elicit empirical expectations, participants are asked to estimate the proportion 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 use a questionnaire similar to the one developed by Krupka and Weber (2013) that evaluates appropriateness as judged by others through a coordination game. Participants rate different behaviors on a scale of 1 (very socially inappropriate) to 5 (very socially appropriate). The questionnaire aims 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 assesses the appropriateness of taking points (extensive margin), and the other question assesses 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 change 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 mea-

sures of positive reciprocity, negative reciprocity, trust, and altruism (Falk et al. (2018)), as well as a reactance scale (Hong and Faedda (1996)), which is a psychological measure associated with the level of conformity to rules and norms.

3 Theory and Hypotheses

This section is divided into three parts: The first part explains that traditional economic theories should predict no differences between the fine and fee conditions. The second part explores potential behavioral changes, emphasizing trade-offs and potential impacts on prosocial concerns, and also discusses that fees and fines may have differing effects. The third part analyzes the channels and investigates social norms as potential mechanisms for influencing behavioral changes.

3.1 Fine vs. Fee

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

Following this perspective, monetary penalties are implemented in various formats and contexts. Environmental regulations often employ a combination of fees and fines to dissuade environmentally harmful actions. Emission permits, typically issued by governmental bodies, function as fees for companies, granting them the privilege to release a specified amount of greenhouse gases. Conversely, companies that violate environmental regulations often face fines as a punitive response.

In general, economic theory only considers the trade-offs associated with these penalties. Different formats may introduce concerns about risk or create a significant time gap between the action and the penalty, which influences behavior. However, given that the underlying trade-offs remain consistent, the specific format should not impact behavior (Tversky and Kahneman (1988)).

Within our experimental framework, we introduce a monetary penalty in two distinct formats: the fine condition, a penalty after the actions, and the fee condition, a penalty before the actions. We design these conditions aiming to maintain consistent trade-offs across both scenarios.

For instance, we intentionally exclude a risk component in the fine condition. By doing so, we ensure that we have the same values in both the fee and fine conditions, enabling a direct comparison of the differences in this small change in the timing of the decision without the influence of risk preferences.

We focus on one essential distinction across fees and fines, the timing of the payment. Whether it occurs before or after. The fine is paid *after* the choice, while the fee introduces the payment *before* the action and introduces a two-stage decision in which the participant has to decide if they are going to take money, and then the amount to be taken. These difference does not directly affect the trade-offs but might influence how individuals react to such penalties.

Given this, fees and fines represent the same fixed cost of 100 points associated with taking money, resulting in an equivalent trade-off. Therefore, the distinction between fees and fines is essentially a framing effect, as it provides no additional information and presents the same set of potential outcomes. Hence, classic economic theory would describe that fee and fine would lead to the same results. This concept is illustrated in our Hypothesis A:

Hypothesis A: There are no differences between fee and fine.

Despite the trade-offs being identical, and classic economic theory predicting no differences across the conditions, fines and fees could potentially lead to distinct behaviors due to the possibility of monetary penalties exerting indirect influences on prosocial motivation. This is elaborated on in the following subsection.

3.2 Fine vs. Fee shaping prosocial behavior

Incentives generally affect the trade-offs associated with a situation, but sometimes incentives can also influence prosocial concerns. For instance, Titmuss et al. (1970) proposed that introducing monetary compensation for blood donation might reduce donations. This hypothesis was tested by Mellström and Johannesson (2008), yielding mixed results, including a decrease in blood donations among female participants when monetary rewards were offered. A similar study by Frey and Oberholzer-Gee (1997) examined support for a nuclear waste storage facility and observed decreased support when monetary compensation was introduced. Gneezy and Rustichini (2000b) demonstrated that offering small monetary rewards led to reduced performance on various tasks, including logical exams. Similarly, Gneezy and Rustichini (2000a) reported that implementing a fine in a daycare

for late-picking parents led to more late pickups.

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

Conversely, rule-following behaviors, as described by Kimbrough and Vostroknutov (2016, 2018), suggest that people have rule-following tendencies even when they are against their monetary interests. For example, participants adhere to red traffic lights in simulations, even when it is costly. In our setting, a monetary penalty could be perceived as a new rule to follow, leading some participants to reduce the amount taken to conform to this new rule or a signal that the behavior is undesirable, potentially causing crowding-in effects and increasing prosocial motivation.

We aim to disentangle the potential shifts and distinctly identify those linked to trade-offs and those related to changes in prosocial concerns. Our design, incorporating twins, is structured to minimize the impact on trade-offs while accentuating potential shifts in prosocial concerns.

To better understand how the twin cases accentuate the changes in prosocial concerns, we illustrate it by analyzing the general impacts of the monetary penalties in our setting. Dictator games are generally analyzed using models of prosocial preferences as in Fehr and Schmidt (1999); Andreoni and Miller (2002) or Charness and Rabin (2002), and we use a simplified inequality aversion model, as in Fehr and Schmidt (1999) in our setting. Consider a 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 ζ captures the level of inequality aversion. The agent's objective is to maximize:

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

In this case, the agent has two options:

1. Indifference to inequality: If $\gamma \leq 0.5$, the agent takes everything, $t^* = y$, and keeps $x + y$.

2. Minimizes inequality: If $\gamma \geq 0.5$, the agent takes enough to keep half, takes $t^* = \frac{(x+y)}{2} - x$, and keeps $\frac{(x+y)}{2}$.

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

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

The agent, facing a penalty, has three options based on different ζ levels[2]:

1. Indifference to inequality: The agent takes everything, $t^* = y$, keeping a total of $(x + y - p)$.

2. Avoids efficiency loss: Due to efficiency loss $(-p)$, takes zero, $t^* = 0$, and keeps the initial endowment, x, avoiding the penalty.

3. Minimizes inequality: The agent takes enough to keep half, $t^* = \frac{(x+y+p)}{2} - x$, redistributing the efficiency loss among participants, taking an extra $\frac{p}{2}$ than the previous case.

The second potential behavioral change illustrates an important aspect of our experiment. To better illustrate it, consider an agent in a $(200, 800)$ scenario. In the control condition, someone with strong inequality aversion takes 300 points, resulting in a 500/500 split. Introducing the penalty, the agent loses 100 points, resulting in a $(100, 800)$ situation. Here, the agent takes 350 points, reaching a 450/450 taking 50 points more than before, a 'more selfish' choice that may be naively seen as a change of prosocial concerns, and a crowding-out effect.

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

Generally, models for prosocial preferences (e.g., Fehr and Schmidt (1999); Andreoni and Miller (2002); Charness and Rabin (2002); Yang, Onderstal, and Schram (2016)) only consider the set of potential outcomes in their utility function, $U(x+t, y-t)$. Given that, these models would predict identical decisions when money is taken in the treatment conditions and the amount taken in the control conditions since they present the same set of potential outcomes, controlling for the trade-off changes:

[2]Different initial endowments would lead to different thresholds. In Appendix 12, we show the threshold for each case in our treatment.

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

Therefore, the observed changes between the control and treatment conditions cannot be attributed to changes in trade-offs; rather, they indicate shifts in prosocial concerns.

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

Models for prosocial preferences (e.g., Fehr and Schmidt (1999) or Andreoni and Miller (2002)) offer similar insights about these agents: Individuals who take large amounts are less likely to stop taking money. When participants take large amounts of money, they show that they prioritize their self-interest over others and, hence, should be less affected by the penalty. Different models and parameters would yield varying thresholds for the amount of money that people would be willing to take and stop after the implementation of the monetary penalty. Hence, it is not possible to fully describe the impact of the trade-offs in such behavior. However, if agents take larger amounts of money in the control condition and stop taking money after the introduction of the penalty, these behavioral shifts can potentially be attributed to crowding-in effects. [3]

We formulate our base hypothesis based on models associated with general trade-off analysis and prosocial preferences. We also assume that we are comparing twin cases, which leads to clear and precise predictions. After each hypothesis, we will explore potential alternative explanations.

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

The introduction of a monetary penalty imposes a fixed cost, which may discourage some agents from taking any points due to the associated efficiency loss. However, if the agent chooses to take money in the treatment condition, they must take the same amount of money as they take in the control condition, as they are facing the same set of possible alternatives when considering the twin cases, as described before. Given these two potential changes, the penalty should reduce the average amount of money taken.

This aggregate change reflects two distinct alterations: the extensive margin, which

[3] In Appendix A, we provide an example of a quadratic function for inequality aversion to illustrate a potential threshold for the amount of money that agents would take in the control condition and stop taking in the treatment condition. In such a case, the maximum amount of money that the individual would take and stop after the introduction of the penalty would be less than 100.

concerns the number of participants taking money, and the intensive margin, which pertains to the amount of money being taken. We highlight those changes before discussing the potential impacts on prosocial preferences.

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

As previously described, when considering the trade-offs and the money loss due to the monetary penalty, some agents may choose to cease taking money due to a fixed cost.

However, as observed in Gneezy and Rustichini (2000a), crowding-out effects might indicate an increase in the number of participants taking money after the penalty is implemented. On the other hand, crowding-in effects and a propensity to follow rules suggest a larger reduction in the number of people taking money.

For instance, the penalty could be perceived as a form of permission to act, reducing the moral concerns associated with the situation. This could lead people to believe that taking money is more socially acceptable, resulting in crowding-out effects. Conversely, if the penalty is perceived as a signal that such behavior is "bad," participants might view taking money as less socially appropriate when the penalty is implemented, leading to more instances of crowding-in effects.

Additionally, the upfront payment of the fee may further influence the moral significance of the decision, a similar argument as Eriksson, Strimling, Andersson, and Lindholm (2017). If this is the case, if the penalty undermines social norms, the fee might lead to higher levels of crowding-out effects than the fine. Conversely, if the penalty highlights prosocial behavior within social norms, the fee might lead to higher levels of crowding-in effects.

We also analyze the individuals who stop taking money when the penalty is implemented. In general, these agents are likely to be those who take small amounts of money, indicating higher levels of prosociality, and the fixed cost associated with the penalty is more likely to affect them.

If agents who stop taking money in the treatment conditions are also taking large amounts in the control condition, it could be evidence of a potential crowding-in effect. As a benchmark, we will compare the amounts taken with the size of the penalty, which is set at 100 points.

Meanwhile, the intensive margin can be described as follows:

Hypothesis 3 - Intensive Margin: If a participant takes points in the treatment condition, there is no difference in the amount taken in the control and treatment conditions.

As previously explained, in the context of twin cases, the set of potential outcomes remains the same, and given that any money is taken, the amount should be consistent.

Crowding-out effects might suggest that people could take money more intensively, while crowding-in effects could indicate that people would take lower amounts.

Similar to the earlier arguments, the concept of entitlement illustrated by Gneezy and Rustichini (2000a) could contribute to a crowding-out effect with fees. Participants might feel they have an even greater right to take money as they already paid to do so, in contrast to fines where the payment occurs simultaneously with the decision. If this is the case, fees could lead to larger crowding-out effects.

3.3 Fine vs. Fee shaping social norms

Social norms have been described as a key component associated with these indirect changes in prosocial concerns, as illustrated by various models and experiments (e.g., Ellingsen and Mohlin (2022); Capraro and Perc (2021); Kimbrough and Vostroknutov (2016); Bénabou and Tirole (2006); Janssen and Mendys-Kamphorst (2004); Gneezy, Meier, and Rey-Biel (2011)). These models vary in aspects related to signaling to others, coordination devices, self-image concerns, or even moral considerations. We focus on one specific way that social norms can indirectly affect behavior: conformity.

The introduction of new incentives might trigger different social norms, similar to what is illustrated by Lane, Nosenzo, and Sonderegger (2023b). Meanwhile, there is an extensive literature describing how individuals conform to social norms (e.g., Bicchieri (2005), Bicchieri (2016), Xiao and Bicchieri (2010)). If the norms shift, and agents conform to these new norms, behavioral changes will occur.

If the monetary penalty leads to "better" social norms, a crowding-in effect can be expected. If the monetary penalty leads to "worse" social norms, a crowding-out effect can be expected. If fees and fines lead to different social norms, it is expected different behavior.

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

Our methodology is also inspired by attribution theory from social psychology (Peterson

et al. (1982); Dykema, Bergbower, Doctora, and Peterson (1996)), examining how individuals perceive causes and motivations behind experiences. This method partially captures the social construction of motivation (entitlement) based on context and individuals.

We illustrate how social norms indirectly affect social concerns with the following utility function: The agent's utility, U, depends on their initial endowment x, the amount taken t, the penalty p, and social norms $N(t, t^k_{\text{emp}}, t^k_{\text{nor}}, t^k_{\text{ent}})$. These norms are integrated into the utility function, where t^k_{emp} represents empirical expectations, t^k_{nor} represents normative expectations, and t^k_{ent} represents perceived entitlement. The norms are context-dependent and each condition, k, can lead to different perceived norms. We also introduce a parameter γ, representing the agent's propensity to conform to norms:

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

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

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

In this scenario, the amount taken, t^*, can be expressed as $t^* = E^k[t] - \frac{1}{2\gamma}$. Thus, higher empirical expectations lead to a larger amount taken. Similarly, if by introducing a penalty, people expect that more money is taken, they are more likely to take more money, leading to crowding-out effects.

Hence, we expect that different conditions would lead to different social norms. We will examine behavioral changes in the extensive and intensive margins. Based on this conformity and these shifts in the social norms, we establish the following hypotheses:

Hypothesis 4 - Norm Shifts: The implementation of monetary penalties impacts social norms (empirical, normative, and entitlement).

Hypothesis 5 - Conformity: Higher empirical/normative/entitlement values for taking any/larger amounts of money are associated with a higher likelihood/larger amounts of taking any money.

If the introduction of the monetary penalty affects social norms/entitlement (Hypothesis 4), and the agent conforms to social norms (Hypotheses 5), we can derive the following corollaries:

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

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

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

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

There are other potential causes for differences between fees and fines. For instance, Zellermayer (1996) describes the pain of payment and suggests that different payment methods might elicit various emotional responses. Similarly, fees and fines can result in similar changes. Read, Loewenstein, Rabin, Keren, and Laibson (2000) explains that people may behave differently when facing narrow or broad bracketing, i.e., when considering problems separately or together. The fee structure creates two decisions, which might influence how agents process information. Such cognitive and emotional aspects can contribute to differences between fees and fines, without necessarily triggering other social norms.

4 Results

The study involved 201 participants, with 101 in the fee condition and 100 in the fine condition, resulting in 4020 decisions. In the first part of the analysis, we describe the behaviors for all the observations, but we primarily focus our main analysis on the twin

cases, with 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 3), resulting in 804 observations for each case.

We checked for order effects, as different sessions started with either control or treatment conditions. We observed no significant difference across the order, as shown in Appendix 14, and thus, all the corresponding treatment sessions are grouped together for data analysis. Appendix 13 demonstrates that the groups are balanced between conditions, with similar age, gender, and ethnicity.

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

4.1 Changes in the prosocial behavior

Aggregate impact:

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

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

We aim to explain the amount taken (*Take*) by individual i in round r. β_0 captures the mean behavior of the control condition in the fee treatment. The variable *Fine* is a dummy for the fine treatment, and β_1 captures the fine treatment effects. *Fee* is a dummy for the fee treatment, and β_2 captures the fee treatment effects. *ControlFine* is a dummy for all sessions in which the participants made decisions on the fine condition, and β_3 captures any potential differences for the control associated with the fine condition and the control condition associated with the fee condition. [4]

We use a random effects model to control for individual differences, and the residuals are clustered at the individual level. After running the regressions, we perform a chi-

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

square test comparing β_1 and β_2 to check if the fee and fine have different impacts.

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 - All data)	(2 - Twin cases)
	Take	Take
Fine	5.426	-6.163
	(6.448)	(7.614)
Fee	-23.35***	-27.78***
	(8.300)	(10.19)
ControlFine	-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) for all observations, *Fine* has no significant impact, and *Fee* leads to a significant reduction in the amount taken; (2) in twin cases, controlling for income effects yields similar results. *ControlFine* shows that there are no differences in the control conditions associated with each treatment.

The results between regression (1) – without controlling for income effects – and regression (2) – controlling for the income effect – are very similar. We consider regression (2) as our primary benchmark. Notably, a statistically significant decrease in the amount taken is observed in the fee condition (-27), providing support for Hypothesis 1. Conversely, the fine condition shows a non-significant decrease (-6). A comparison of the fee and fine treatment impacts reveals a marginally significant difference ($\chi^2(1) = 2.89, p = 0.0894$), hence the fee leads to a marginally bigger impact than the fine. To illustrate this difference, we can observe Figure 1:

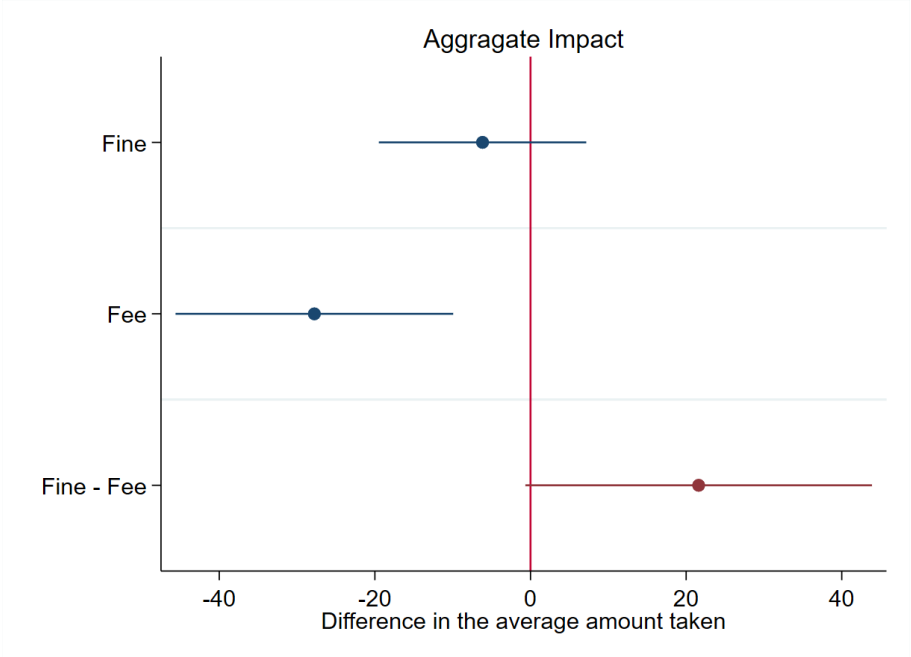

Figure 1: Treatment effects for the twin cases at the aggregate level: the changes in the amounts taken for each condition, Fee and Fine, and the differences between them (Fine-Fine) are presented with 95% confidence intervals.

Result 1: *At the aggregate level, the introduction of a* ***fee*** *results in a significant reduction in the amount taken compared to the scenario with no penalty, whereas the introduction of a* ***fine*** *does not lead to a significant change.*

Result 1A: *At the aggregate level, there are marginally significant differences in the treatment effects of the* ***fee*** *and the* ***fine****, with the* ***fee*** *causing a significantly greater reduction in the amount taken compared to the fine, when compared to situations with no monetary penalty.*

To gain a deeper understanding of these behavioral shifts, we can examine the amount of money taken in each condition and each case, as observed in Table 4

		Fine			Fee			
Twin	**Case**	**Control Amount Taken**	**Treatment Amount Taken**	**Diff**	**Control Amount Taken**	**Treatment Amount Taken**	**Diff**	**Diff-in-Diff**
1	(100,800)	505.35	525.54	20.19 [0.101]	514	518.9	4.9 [0.78]	15.29 [0.47]
	(200,800)	470.79	513.83	42.37*** [0.00]	486.4	509.4	23 [0.148]	19.37 [0.35]
2	(170,730)	450.69	455.14	4.45 [0.74]	450.1	452.9	2.8 [0.87]	1.65 [0.94]
	(270,730)	414.45	447.82	33.36** [0.02]	435.2	445.2	10 [0.59]	23.36 [0.33]
Decoy 1	(360, 510)	274.75	270.89	-3.86 [0.80]	240.2	238.6	-1.6 [0.90]	-2.26 [0.91]
3	(500,400)	153.46	148.11	-5.34 [0.67]	161.7	104.9	-56.8*** [0.00]	51.45*** [0.00]
	(600,400)	152.67	138.01	-14.65 [0.18]	173.8	112.9	-60.9*** [0.00]	46.24*** [0.00]
4	(550,350)	139.50	134.15	-5.34 [0.59]	143.7	83.5	-60.2*** [0.00]	54.85*** [0.00]
	(650,350)	135.74	125.34	-10.39 [0.28]	146.1	90.9	-55.2*** [0.00]	44.80*** [0.00]
Decoy 2	(620,310)	81.98	75.44	-6.53 [0.29]	81.1	41.6	-39.5*** [0.00]	32.96*** [0.00]

p-values in brackets referenced to a random effect model with standard error cluster on the individual level

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

Table 4: **Amount taken by each case and condition:** The average amount taken in each case given and their respective conditions and treatments, along with their differences. The last column describes the differences-in-differences across fee and fine treatment effects.

As the initial endowment increases, the available amount to be taken decreases, illustrating the decision-making dynamics[5]. In the fine condition, when the agent starts with more money than their opponent, there is a consistent increase in the amount taken, reaching statistical significance in some instances. Conversely, when the agent begins with less money, the fine leads to a systematic reduction, although this effect does not reach statistical significance. In contrast, in the fee condition, the fee results in nonsignificant increases when the agent starts with less money but leads to systematic and statistically significant decreases when the agent has more money. Hence, it seems that the aggregate impact of the fee is driven by cases in which the agent starts with more money.

To gain a deeper understanding of these differences, we analyze the impact of both the extensive margin, i.e., the number of instances in which money is taken, and the intensive margin, i.e., the amount of money taken when money is taken.

Extensive margin:

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 5 presents the results, with Regression (3) using the entire dataset, and Regression (4) focusing on the twin cases.

[5]Further details are provided in the Appendix 15

	(3 - All data) Participation	(4 - Twin cases) Participation
Fine	-0.514***	-0.388**
	(0.139)	(0.156)
Fee	-1.269***	-0.962***
	(0.159)	(0.159)
ControlFine	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 5: **Extensive margin:** Treatment effects on the number of times participants took any money: (3) For all observations, both *Fine* and *Fee* result in a significant reduction in participants. (4) In twin cases, controlling for income effects yields similar results. *ControlFine* shows that there are no differences in the control conditions associated with each treatment.

The observations provide evidence supporting Hypothesis 2 for both regression (3) and (4). Using regression (4) as our main benchmark, there is a decrease in the percentage of cases where points are taken in both the fee and fine conditions. Translating the logit differences into numbers, we observe a reduction from 80.19% to 64.64% for the fee condition and from 80.65% to 75.06% in the fine condition, accounting for the twin cases.

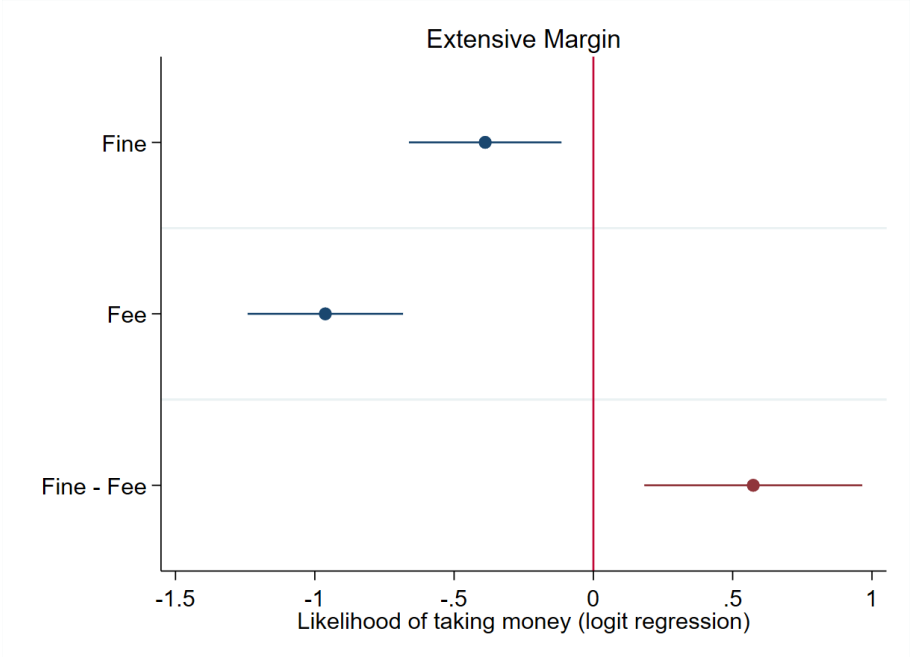

Figure 2: Treatment effects for the twin cases on the extensive margin: the changes in the instances in which money is taken for each condition, Fee and Fine, and the differences between them (Fine-Fine) are presented with 95% confidence intervals.

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

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

Result 2: *At the extensive margin, implementing both the* ***fee*** *and the* ***fine*** *significantly reduces the number of cases where money is taken, compared to the scenario with no monetary penalty.*

Result 2A: *At the extensive margin, there are significant differences in the treatment effects of the* ***fee*** *and the* ***fine****, with the* ***fee*** *causing a significantly greater reduction in the number of instances that money is taken compared to the fine, when compared to situations with no monetary penalty.*

We can also check the participation for each case and condition, similar to what the previous analysis, and summarized at Table 6:

		Fine			Fee			
Twin	**Case**	**Control Participation**	**Treatment Participation**	**Diff**	**Control Participation**	**Treatment Participation**	**Diff**	**Diff-in-Diff**
1	(100,800)	1	0.98	-0.02	0.99	0.93	-0.06**	0.04
				[0.156]			[0.031]	[0.318]
	(200,800)	1	0.98	-0.02	0.99	0.95	-0.04	0.02
				[0.156]			[0.01]	[0.471]
2	(170,730)	1	0.98	-0.02	0.99	0.93	-0.06**	0.04
				[0.156]			[0.031]	[0.318]
	(270,730)	1	0.97	-0.03*	0.99	0.94	-0.05*	0.02
				[0.081]			[0.056]	[0.515]
Decoy 1	(360, 510)	1	0.84	-0.16***	0.99	0.72	-0.27***	-0.11*
				[0.000]			[0.000]	[0.061]
3	(500,400)	0.58	0.50	-0.08*	0.6	0.31	-0.29***	0.21**
				[0.071]			[0.000]	[0.002]
	(600,400)	0.57	0.50	-0.07*	0.64	0.36	-0.28***	0.21**
				[0.087]			[0.000]	[0.001]
4	(550,350)	0.58	0.50	-0.08**	0.64	0.3	-0.34***	0.25***
				[0.047]			[0.000]	[0.000]
	(650,350)	0.60	0.52	-0.08*	0.65	0.33	-0.32***	0.24***
				[0.071]			[0.000]	[0.000]
Decoy 2	(620,310)	0.59	0.44	-0.15***	0.58	0.36	-0.35***	0.20**
				[0.000]			[0.000]	[0.002]

p-values in brackets referenced to a random effect model with standard error cluster on the individual level

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

Table 6: **Participation by each case and condition:** Average number of instances in which money has been taken in each case given and their respective conditions and treatments, along with their differences. The last column describes the differences-in-differences across fee and fine treatment effects.

When the agent starts with less money than their opponent, they consistently take money. The introduction of the fine results in a small, non-significant reduction in the number of cases where money is taken. In contrast, the introduction of the fee leads to a significant reduction, although not significantly different from the fine.

When the agent starts with more money than their opponent, a significant share of participants already refrains from taking money. The fine consistently produces a marginally significant reduction in the number of cases where money is taken. However, the fee has a drastically significant impact, leading to even larger reductions, significantly surpassing the effect of the fine.

We also investigate the behavior of the agents who stop taking money during the control condition, i.e., how much they are taking in the control condition for the twin case in which they stop taking money in the treatment condition. Figure 3 displays the distribution of the amount taken for the same respective cases.

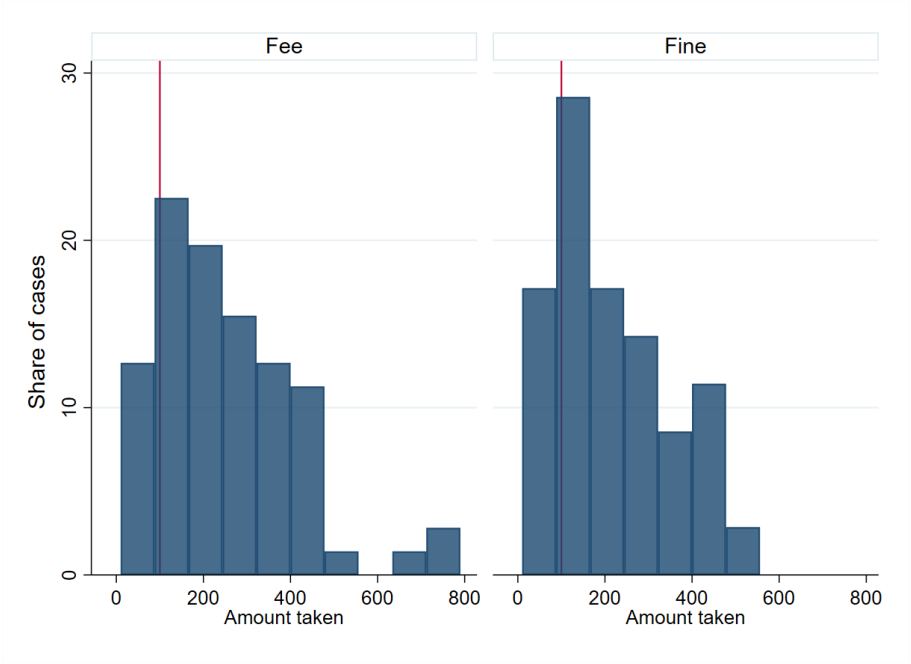

Figure 3: The distribution of the amount taken among those who did not take money in the treatment conditions. On the left side, the amount taken in the control condition by those who did not take money in the fee treatment. On the right side, the same information is presented for the fine treatment.

Participants consistently take more than 100 points. The fee results in an average reduction of 248 points, whereas the fine condition shows a reduction of 200 points, with no significant differences between the treatment conditions ($\chi^2(1) = 0.88, p = 0.3482$). In

approximately 50% of the cases, participants take more than 200 points, and in around 30% of the cases, they take more than 300 points but then cease taking money in the treatment conditions. As a benchmark criterion, we compare the amount taken with the 100-point cost of the monetary penalty, and the average amount taken is significantly different ($\chi^2(1) = 42.50, p = 0.0000$).

As the range of amounts that can be taken changes across the conditions, we can also observe the share kept by the dictator - (*Take* + *Initial Endowment for dictator*) / (*sum of initial endowments*) to create the same unit across all cases. Figure 4 shows the distribution of these values.

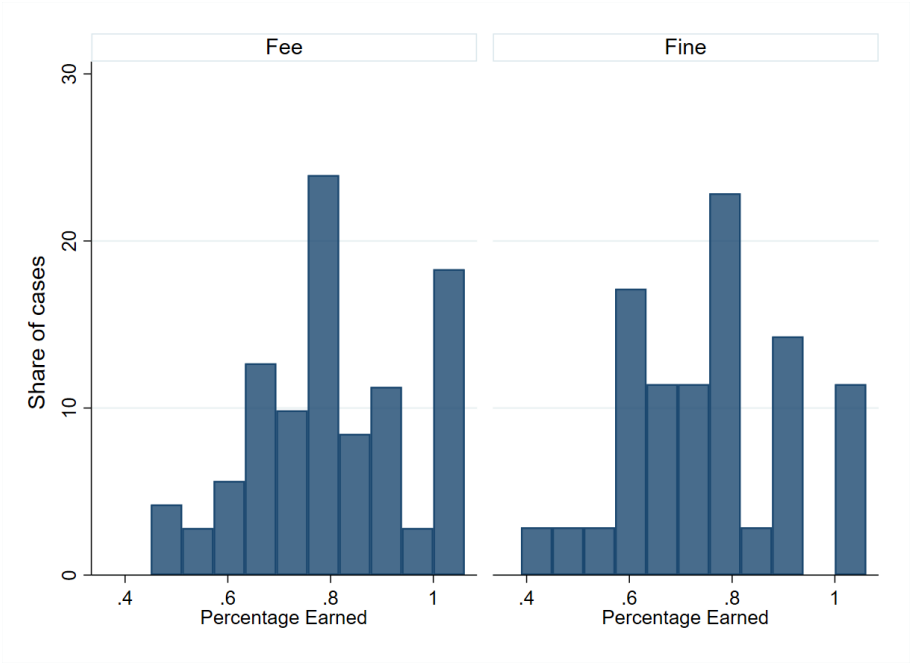

Figure 4: The distribution of the total share kept among those who did not take money in the treatment condition is shown on the left side. On the left side, the share kept in the control condition by those who did not take money in the fee treatment is displayed, while on the right side, the same information is presented for the fine treatment.

On average, dictators obtain around 80% and 77% of the total available in the fee and fine conditions, respectively, for their specific control conditions and then stop taking any money. In some cases, these ratios are extremely high. For example, in the control condition associated with the fee condition, dictators obtain 100% of the money in 18.3% of cases, while in control conditions associated with the fine condition, this occurs in 11.43% of cases, and these individuals decide to stop taking any money after the penalty is imposed. These substantial reductions in the amount taken provide evidence

of a crowding-in effect.

Intensive margin:

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

To control for this aspect, we specifically select the cases where money was taken in the treatment condition and match those cases with the same cases for the same participants in their respective control conditions, ensuring consistency across participants and cases in the regression. Regression (6) presents the results when we paired with the same case, while regression (7) pairs with their twin case, controlling for individual and income effects.

Notice that the coefficient, *ControlFine*, is intended to capture whether the participants who are willing to take money after the fee or fine conditions significantly differ. If this is the case, *ControlFine* will account for these differences. Table 7 offers additional details:

	(5 - All data)	(6 - Only same participants)	(7 - Twin cases)
	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)
ControlFine	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 7: Intensive margin: treatment effects on the amount of money taken are examined, conditional on taking money in the treatment condition. Case (5) includes all instances where money is taken, while case (6) pairs the exact same cases for the control and treatment conditions within the same individual. Case (7) pairs cases with their twins, also controlling for income effects. Across all regressions, *Fee* and *Fine* lead to significant increases in the amount taken. Notably, *ControlFine* reveals some differences across individuals who persist in taking money given a fine or a fee, although these differences do not appear robust across the regressions.

The results contradict hypothesis 3, suggesting increases in the amount taken, and we observe crowding-out effects for all regressions. After controlling for income effects, regression (7), both the fee and fine conditions lead to a significant increase in the amount taken - 15.45 and 25.31, fine and fee respectively. We conducted a chi-square test to compare the fee and fine treatment effects ($\chi^2(1) = 0.73, p = 0.3933$), revealing no significant differences between them. The results can be observed in figure 5:

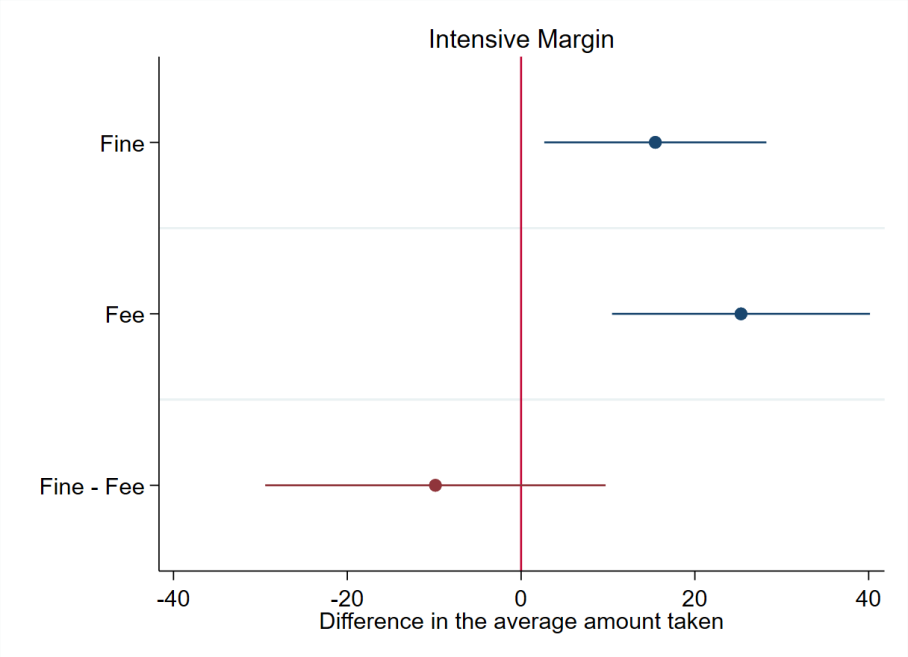

Figure 5: Treatment effects for the twin cases at the intensive margin – the changes in the amounts taken for each condition, given that money was taken in the treatment condition (fee and fine), and the differences between them (Fine-Fee), are presented with 95% confidence intervals. The fine exhibits a significant average increase of 15, while the fee significantly reduces by 25. The difference of 10 between the fee and fine was not significant.

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

Result 3: *At the intensive margin, both the* ***fee*** *and the* ***fine*** *lead to an increase in the amount taken compared to the cases with no monetary penalty.*

Result 3A: *At the intensive margin, there are no significant differences between the treatment effects of* ***fee*** *and* ***fine****.*

We can examine variations in the intensive margin for each case and condition, similar to the previous analysis, and present a summary in Table 8:

		Fine			Fee			
Twin	**Case**	**Control Participation**	**Treatment Participation**	**Diff**	**Control Participation**	**Treatment Participation**	**Diff**	**Diff-in-Diff**
1	(100,800)	521.63	560.34	38.70**	506.08	536.28	30.20**	-8.50
				[0.001]			[0.003]	[0.590]
	(200,800)	489.12	536.07	46.94***	470.65	521.96	51.31***	4.36
				[0.000]			[0.000]	[0.800]
2	(170,730)	449.41	490.70	41.29	448.13	464.90	16.76	-24.52*
				[0.156]			[0.109]	[0.080]
	(270,730)	431.59	475.74	44.14**	409.25	457.21	47.95**	3.81
				[0.001]			[0.00]	[0.834]
Decoy 1	(360, 510)	247.39	291.42	44.027**	257.03	301.86	44.82***	0.79
				[0.001]			[0.000]	[0.967]
3	(500,400)	144.92	189.76	44.83*	157.31	203.98	46.666**	1.82
				[0.071]			[0.004]	[0.944]
	(600,400)	174.52	189.24	14.72	165.30	188.44	23.13**	8.41
				[0.171]			[0.075]	[0.618]
4	(550,350)	135.75	149.08	13.33	132.66	168.22	35.55**	22.22
				[0.212]			[0.001]	[0.146]
	(650,350)	146.15	156.46	10.30	128.82	154.10	25.28**	14.97
				[0.466]			[0.031]	[0.41]
Decoy 2	(620,310)	1.80	21.37	19.56	46.99	71.88	24.88**	5.32
				[0.066]			[0.002]	[0.689]

p-values in brackets referenced to a random effect model with standard error cluster on the individual level

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

Table 8: **Amount taken by each case and condition when money was taken in the treatment condition:** The average amount taken by individuals who consistently took money in the treatment condition for each case, their respective conditions and treatments, and the resulting differences. The last column illustrates the differences-in-differences between fee and fine treatment effects.

Considering the decisions of participants who kept taking money in the treatment condition, both Fee and Fine demonstrate systematic increases in the amount taken when implemented compared to their respective controls. The increase in the fine condition is consistently significant only when the agent starts with less money than their opponent, whereas the fee consistently increases values similarly across all cases. However, the differences are not significant.

In summary, our findings highlight the significant and heterogeneous impacts of introducing monetary penalties on prosocial behavior, with noteworthy distinctions between the fee and fine conditions. Some participants become less likely to take money after the penalty's introduction, even if they had previously taken substantial amounts, indicating a crowding-in effect. Conversely, among participants who persist in taking money despite the penalty, they do so more intensively, demonstrating a crowding-out effect.

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

We also observed differences in the impacts across various cases. The analysis delving into the relationship between inequality and behavioral changes is discussed and illustrated in Appendix D. In general, when the agent starts with more money than the opponent and is compared with their respective control conditions, the fee condition leads to further decreases in the instances of money being taken compared to the fine condition. However, for agents consistently taking money, the amount taken systematically increases when the penalty is implemented, more regularly and consistently in the fee condition compared to the fine condition. Hence, it appears that starting with more money enhances the crowding-in and crowding-out effects of the fee condition.

4.2 Social Norms and Entitlement

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

We hypothesize that monetary penalties trigger shifts in social norms, and there is a positive monotonic relationship between these behaviors and norms/entitlement. This suggests that if something is perceived as more expected, appropriate, or entitled, individuals are more likely to behave accordingly. To test this, we first examine whether the introduction of a monetary penalty affects these measures of social norms and entitlement.

Then, we assess whether the observed behavioral changes can be linked to these potential norm shifts.

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

The first aspect reflects the extensive margin: For empirical expectations, we ask the participants to consider 100 other participants and inquire about how many would take money. For normative expectations, we inquire about the perceived appropriateness levels for others taking any amount, and for perceived entitlement, we ask participants about their perception of how entitled others were to take any amount

To analyze the potential changes, we employ the same regression as in the previous question, adjusting the dependent variable for each measure of social norm/entitlement. The regressions are illustrated in Table 9, 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)
ControlFine	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 9: Social norms for the extensive margin—empirical expectations (8), normative expectations (9), and perceived entitlement (10).

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

Result 4 - Norm Shifts (Extensive Margin): *Both the fee and the fine result in*

significant shifts in social norms associated with the extensive margin. Participants anticipate fewer people taking money and attribute lower scores to normative and entitlement levels for taking any money when a penalty is present compared to a situation with no penalty. No significant difference in the impact of the fee and fine is observed.

The second aspect is associated with the intensive margin. For empirical expectations, we inquire about the average amount of money taken by the same 100 participants. To better proxy the intensive margin, we weight this value by the expected number of participants taking money, from the previous question, yielding the regular intensive margin. For normative and entitlement aspects, we asked participants to express the appropriateness/perceived entitlement for taking approximately 70% of the total amount.

The regressions are presented in Table 10. Regression (11) outlines a linear regression with random effects for empirical expectations, while regressions (12) depict the weighted empirical expectations. Regressions (13) and (14) present analyses for normative expectations and entitlement.

As we aim to examine the impact on the intensive margin and capture the crowding-out effect, we assess norm changes for those agents who continue taking money in the treatment condition. In other words, we analyze the norm change for the sample used in the previous intensive margin analysis.[6]

[6]For the Weighted Empirical Expectations, in a few cases, participants anticipated that no one would take money, preventing the creation of 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)
ControlFine	-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 10: Social norms for the intensive margin: (10) empirical expectations, (11) weighted empirical expectations, (12) normative expectations, and (13) perceived entitlement. Fee and Fine conditions manifest distinct changes across these measures, with some being significant and others not.

The penalties do not induce changes in general empirical expectations regarding the amount taken, even though people expect an increase. However, weighted empirical expectations display a significant increase in the fine condition and a substantial increase in the fee condition, though not deemed significant due to the high variance associated with this new measure.

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

Result 4 - Norm Shifts (Intensive Margin): *Both the fee and fine result in significant shifts in social norms related to the intensive margin. Participants assign higher scores to normative levels for taking any money when a penalty is present compared to the situation with no penalty. Additionally, the fee leads to higher entitlement scores than its respective control condition.*

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

norms/entitlement could potentially explain behavioral changes.

We leverage behavioral observations from the four cases where we have measured social norms/entitlements to replicate the earlier findings. Subsequently, we conduct two new regressions: one to explore the new treatment effects after incorporating social norms/entitlement, and the other regression introduces an interaction term between social norms and the treatments.

Specifically, we start by replicating the results previous results using only the cases in which the norms were measures (twin 2 & 3):

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

Subsequently, we conduct the following regression:

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

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

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

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

We also use the following regression:

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

This regression adds an interaction term between the *Fee* dummy that captures the treatment condition, and each social norm. By doing so, we can analyze if the social norms affect the fee and the fine differently.

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 Table 11, regression (15) aims to replicate the previous results for the extensive margin using a smaller selected sample (2 twin cases where norms were measured) through linear regression[7]. In regression (16), we incorporate social norms/entitlement into the regression. In regression (17), interaction terms are also added. Regressions (18), (19), and (20) reproduce the same results for the intensive margin (Take) using linear regression.

[7]To facilitate the comparison of coefficients across regressions.

	(15) Participation	(16) Participation	(17) Participation	(18) Take	(19) Take	(20) 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)
ControlFine	-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 11: Channels: replicate extensive margin regressions (15), incorporate social norms (16), and introduce interaction terms (17). Replicate intensive margin regressions (18), include social norms (19), and add interaction terms (20). The results are robust for the interaction terms, except for endowment which is not robust. The social norms capture the treatment effects and are positively correlated with behavior.

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

Secondly, the coefficients for social norms and entitlement are positive and significant

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

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

When comparing the results of regression (15) and (16) to analyze 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 social norms partially account for the treatment effects for the extensive margin. However, the gap between fee and fine conditions remains similar even when controlling for social norms.

When comparing the results of regressions (18) and (19) to analyze 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.

The regressions remain fairly consistent when the interaction terms are added, comparing regression (16) and (17), and regressions (19) and (20). The only divergence is observed for the impacts of *Entitlement*, which is not robust across the equations. This suggests that both empirical and normative expectations play similar roles for fee and fine, while entitlement does not.

Result 5: 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 5A: 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 compare the impact of stylized fines and fees on behavior, aiming to identify differences in effectiveness and pinpoint the effects of monetary penalties on prosocial concerns. We systematically observe differences across both fees and fines and exemplify the heterogeneous impacts of monetary penalties on behavior: Some agents exhibit crowding-out effects (e.g., Gneezy and Rustichini (2000a)), indicating reduced prosocial motivation after implementing the penalty, while others display crowding-in effects (e.g., Kimbrough and Vostroknutov (2016)), boosting prosocial motivation due to the penalty. Moreover, we observe that the implementation of monetary penalties induces shifts in social norms, and these shifts can partially explain the penalties' impacts on prosocial concerns. Our study provides a nuanced understanding of the effects of monetary penalties on human behavior, offering insights into policy considerations regarding how different penalty formats can lead to divergent outcomes.

We employ dictator games, allowing individuals to take money from others, and introduce fees or fines aiming to reduce the amount of money being taken. We design our experiment penalties to avoid other confounding factors, focusing on a crucial difference across fees and fines: Fees impose penalties *before* the action. In our case, it establishes a first-stage decision where participants must choose whether to take money. Conversely,

fines are deducted *after* any money is taken. Both fees and fines result in the same trade-offs, consistently tied to a fixed cost of 100 points for taking money. Therefore, any observed behavioral changes should be regarded as framing effects, and classic economic theory suggests no differences across the conditions (Tversky and Kahneman (1988)). However, as observed, they might affect prosocial concerns differently.

To gain a comprehensive understanding of the impact of monetary penalties on prosocial concerns, we structure our dictator games into multiple situations, creating paired twin cases. These pairs enable us to control for potential income effects linked to the penalty. Consequently, any observed behavioral changes in these comparisons should not be solely attributed to trade-off analysis but rather associated with the indirect impacts of the penalty on prosocial behavior.

When the penalty is introduced, many participants refrain from taking money, even if they had previously taken large amounts, leading to changes in the extensive margin. The fee treatment leads to a roughly 15% reduction in the number of cases where money is taken, while the fine treatment results in a 5% reduction. This difference is significant, indicating that people were acting more prosocially in the fee condition, showing a stronger crowding-in effect.

Moreover, a significant portion of these participants exhibited substantial money-taking behavior, with approximately 50% taking more than 200 points, and roughly 15% taking the entirety of the available money. Noteworthy is the absence of major distinctions between the fee and fine conditions. Disentangling the direct or indirect impacts on the extensive margin poses a challenge, as we lack control over the income effect associated with the decision to take or abstain from taking money. Diverse models (e.g., Fehr and Schmidt (1999); Andreoni and Miller (2002); Charness and Rabin (2002)) and various parameters would propose distinct thresholds for the amount of money that could be taken without incurring a penalty and what the penalty would deter. However, these significant reductions, especially among individuals taking everything, cannot be solely attributed to trade-off factors. Individuals who take everything exhibit low prosocial concerns and should, in theory, be proportionally less affected by the penalty. Consequently, these substantial reductions provide evidence of individuals becoming more prosocial, indicating a rule-following tendency akin to Kimbrough and Vostroknutov (2016) and suggesting a potential crowding-in effect. Importantly, no significant difference was observed among these individuals across fee and fine conditions.

Participants who persist in taking money after the penalty's implementation exhibit a noteworthy escalation in the amount they acquire, observed in both the fee and fine

conditions, reflecting changes in the intensive margin. This increase persists even when adjusting for income effects and comparing decisions within the same individual. The findings suggest that the penalty serves as a motivator for these individuals to act more selfishly, engaging in a more intensive pursuit of money, indicative of a crowding-out effect.

We observe no significant differences between fees and fines, but there are some small and non-robust variations in the type of individuals who continue taking money and the distribution of the amount of money taken across conditions. It is possible that the fee condition could lead to relatively larger crowding-out effects, but our setting is not able to capture that. Hence, based on our results, the crowding-out effects are roughly the same across fees and fines.

The impact at the aggregate level reflects the combined effects on the intensive and extensive margins. The fine was inefficient and showed no significant impact, as the intensity of the amount of money being taken by those who continued to take compensates for the reduction associated with the lower number of people taking money. Since the crowding-out effects are roughly the same across conditions but the fee condition leads to bigger crowding-in effects, the fee condition significantly reduces the total amount taken.

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

Our findings highlight that the format of the penalty can result in different impacts, with fees proving more effective than fines. This echoes discussions by Bicchieri and Dimant (2019) and Bowles (2016), emphasizing the need for careful consideration in interventions, as the message and format can yield diverse outcomes.

Concerning monetary penalties, our results reveal that subtle changes in the setting can lead to diverse outcomes, providing potential insights for future research and interventions. This observation extends to domains like environmental legislation, which employs fines to deter environmental damage, alongside the establishment of licenses (similar to fees) to permit specific behaviors. Contemporary approaches, such as carbon markets, might bring about even more moral changes, as described by Falk and Szech (2013). Despite appearing similar in economic theory, these different setups can elicit distinct behavioral responses and trigger different moral perceptions. Our results emphasize the importance of analyzing the moral impacts of each setting to create interventions that are truly

effective.

Moreover, our study also demonstrates how the literature can provide evidence for both crowding-out effects, as seen in Gneezy and Rustichini (2000a), and crowding-in effects, as observed in Kimbrough and Vostroknutov (2016). If agents are rule-followers, they should not lead to crowding-out effects; if their prosocial concerns deteriorate, they should not be rule-followers. We show that the impacts are actually heterogeneous, and the balance between different participants leads to the overall results. This is further exemplified when we observe that inequality also plays a role in the size of each impact, with crowding-in effects in the fee condition being systematically higher.

As the monetary penalty affects individuals differently, the overall impact changes according to the situation, the design, and the balance across the heterogeneous groups—those who use the penalty as a reason to cooperate and those who use the penalty to justify their actions. Further research should try to investigate who and in which contexts individuals are more likely to follow crowding-out or crowding-in, leading to significant insights for better legislation and interventions

Our study investigates potential mechanisms behind behavioral changes. We focus on social norms and perceived entitlement as potential explanations for behavioral changes.

As shown by Lane et al. (2023b), the implementation of laws can influence social norms. Previous literature demonstrates that people conform to social norms (e.g., Xiao and Bicchieri (2010); Krupka and Weber (2013); Bicchieri (2005)). The penalty might trigger different norms, leading to varied behaviors. Moreover, social norms have been identified as plausible mechanisms behind the crowding-out and crowding-in effects, as described by Ellingsen and Mohlin (2022); Kimbrough and Vostroknutov (2016) and Gneezy et al. (2011). We also explore the role of entitlement in behavior as entitlement is often cited as a potential explanation for the crowding-out effect (e.g., Gneezy and Rustichini (2000a); Bénabou and Tirole (2006)), as paying the penalty might make individuals feel like they have the right to do so.

Regarding social norms, we categorize them into empirical and normative expectations, following the terminology of Bicchieri (2005, 2016), and use Krupka and Weber (2013)'s method to measure normative expectations. To measure entitlement, we developed a new methodology by adapting Krupka and Weber (2013), using a coordination game to incentivize participants to consider the group's opinion.

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

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

This aligns with Lane et al. (2023b), showing how the implementation of a monetary penalty can also trigger different social norms. However, it also shows differences in the extensive and intensive impact of the penalties, which was not analyzed by Lane et al. (2023b). Moreover, the results corroborate findings by Ellingsen, Johannesson, Mollerstrom, and Munkhammar (2012) and Eriksson et al. (2017), demonstrating that framing effects can influence expectations, and then change behavior.

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

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

Our results offer a direct examination of the role of norms in crowding-out and crowding-in effects, highlighting the influence of social norms on indirect changes in prosocial concerns. This resonates with the insights discussed by Kimbrough and Vostroknutov (2016), underscoring the impact of social norms on social preferences. Unlike perspectives presented by Bénabou and Tirole (2006) or Janssen and Mendys-Kamphorst (2004), which suggest that norms influence behavior through signaling to others, our findings exemplify how norms affect behavior through conformity.

The observed changes in prosocial concerns in our setting are individual, with agents altering their personal evaluations of the situation independently. This underscores models like Krupka and Weber (2013) or Akerlof and Kranton (2000), which directly incorporate social norms into the utility function, and models such as Ellingsen and Mohlin (2022) or Capraro and Perc (2021), which aim to capture similar changes with considerations of duties or morals. Consequently, our results indicate that changes in prosocial concerns can be explained by shifts in norms and conformity.

In alignment with studies like Gerlach and Jaeger (2016); Eriksson et al. (2017), our results echo the findings of Ellingsen et al. (2012). Their work demonstrated that framing

effects create different expectations in game-theoretical settings, influencing behaviors. Similarly, our experimental setup maintains identical trade-offs and choices, varying only the decision format. In parallel, behaviors adapt to align with new perceived expectations and social norms. The implementation of incentives has the potential to reshape individuals' perceptions of a situation, resulting in substantial behavior changes, even when alterations in trade-offs may seem minimal.

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

Our new methodology for capturing endowment also has interesting implications. This approach aligns with attribution theory in social psychology, as discussed in Peterson et al. (1982) and Dykema et al. (1996), which explores how individuals perceive the causes and motivations behind everyday experiences. People act differently given different motivations; for example, the same action might have different reactions depending on whether it is perceived as coming from a malicious intention or not.

This methodology has interesting implications and can be expanded to capture additional aspects related to concerns about social image and motivated reasoning, as explored in previous studies such as Epley and Gilovich (2016). Further research could extend these methods to explore different motivations attributed to various behaviors. For example, Fischer and Teixeira (2023) employs a similar methodology to examine how different motivations are attributed to males or females for the same behavior, shedding light on gender differences and their underlying causes. Analyzing these motivations can enhance our understanding of norms, cognition, and decision-making.

Our study not only enhances our understanding of the impact of monetary penalties on behavior but also explores the heterogeneous effects of these penalties, highlighting distinctions between fees and fines. Through a thorough analysis of the motivations driving these behavioral changes, we contribute to a more comprehensive comprehension

of how financial incentives and deterrents influence individual decision-making. These insights provide valuable guidance for future research and offer the potential to inform the design of more effective policy interventions.

6 Conclusion

Monetary penalties are a common tool for discouraging undesirable behavior, yet their precise impact is not clear, as they can have indirect effects on prosocial concerns, leading to unexpected results. Such indirect impacts might even make some penalty settings more effective than others, even though they reflect the same trade-offs.

We use a modified dictator game in which participants can take money from others and implement a penalty in two formats: a "fine" imposed after taking money and a "fee" paid before taking money. Our findings reveal systematic differences between fines and fees. Moreover, we demonstrate that monetary penalties have indirect and heterogeneous impacts on individuals. For many, the penalty serves as motivation to stop taking money, even when they were previously engaging in it intensely, illustrating crowding-in effects—an increase in prosocial concerns. On the other hand, some individuals take more money after the introduction of the penalty, demonstrating a crowding-out effect—a decrease in prosocial concerns.

The interplay of these forces is context-dependent, as exemplified by the different impacts observed with fees and fines. In our study, fines exhibit a balance of these heterogeneous effects and produce no significant aggregate impact. However, when the same penalty is introduced as a fee, it proves effective, with the crowding-in effect dominating the crowding-out effect, resulting in fewer instances in which money is taken and a lower aggregate amount taken, compared to the fine.

Furthermore, our observations indicate that the introduction of monetary penalties shifts perceived social norms. For example, people believe that taking any amount of money is less socially appropriate when the penalty is implemented compared to no penalty, but they also believe that taking larger amounts of money is more socially appropriate when the penalty is implemented compared to no penalty. These shifts in social norms can partially account for behavioral changes, providing an explanation for both crowding-in and crowding-out effects.

In summary, we demonstrate that monetary penalties have heterogeneous impacts on behavior, leading to crowding-out effects and crowding-in effects for different individuals. The format of the penalty affects these effects, with significant differences depending on

whether the penalty is paid after (as in a fine) or before (as in a fee). The implementation of monetary penalties not only affects the trade-off but also shifts the perceived norms associated with the situation. These changes in norms can partially explain the crowding-out and crowding-in effects.

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Gerlach, P., & Jaeger, B. (2016). Another frame, another game? explaining framing effects in economic games.

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.

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.

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

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

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

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

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

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

Tversky, A., & Kahneman, D. (1988). Rational choice and the framing of decisions. *Decision making: Descriptive, normative, and prescriptive interactions*, 167–192.

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

ζ and the thresholds

		High prosocial concern		Low prosocial concern	
Case	**Control**	$\zeta \geq 0.5$	**C>Treat**	$\zeta \leq 0.5$	**C>Treat**
(100/800)	$100 - \zeta(700)$	400	$\zeta < -\frac{3}{7}$	$800 - 800\zeta$	$\zeta > 1$
(170/730)	$170 - \zeta(560)$	400	$\zeta < -\frac{23}{56}$	$800 - 800\zeta$	$\zeta > \frac{23}{8}$
(200/800)	$200 - \zeta(600)$	450	$\zeta < -\frac{5}{12}$	$900 - 900\zeta$	$\zeta > \frac{7}{3}$
(270/730)	$270 - \zeta(460)$	450	$\zeta < -\frac{9}{23}$	$900 - 900\zeta$	$\zeta > \frac{63}{47}$
(500/400)	$500 - \zeta(100)$	400*	$\zeta < 1$	$800 - 800\zeta$	$\zeta > \frac{3}{7}$
(550/350)	$550 - \zeta(200)$	400*	$\zeta < \frac{3}{4}$	$800 - 800\zeta$	$\zeta > \frac{5}{12}$
(600/400)	$600 - \zeta(200)$	450*	$\zeta < \frac{3}{4}$	$900 - 900\zeta$	$\zeta > \frac{3}{7}$
(650/350)	$650 - \zeta(300)$	450*	$\zeta < \frac{2}{3}$	$900 - 900\zeta$	$\zeta > \frac{5}{12}$

Table 12: ζ and the respective threshold for stopping taking money after the introduction of the penalty are presented for each case. The first column represents the cases, the second column shows the utility in case the agent does not take money. The third column describes that if $\zeta > 0.5$, the agent takes half of the amount available, keeping such value as utility. The fourth column compares the control with half of the amount taken, creating the threshold for ζ that would make the agent cease taking money. An asterisk (*) represents the choice that would be possible that would be possible if the agent could give money. However, anyone with $\zeta > 0.5$ starting with more money would never take money, as the money taken would increase the inequality. The fifth column describes agents with $\zeta \geq 0.5$, who would take everything. The last column describes the thresholds for ζ that would lead to moving to take zero.

The table describes situations in which the agent would take money and cease taking any money using an inequality aversion model.

Regarding inequality aversion, two possibilities exist. The agent may exhibit high prosocial concern, indicated by γ exceeding 0.5, leading the agent to claim half of the total available to rectify inequality. Alternatively, the agent may have low prosocial concern, as indicated by γ falling below 0.5, prompting the agent to seize all available resources.

In situations where the agent has high prosocial concerns, there is no circumstance in which the agent is willing to intermittently take and stop taking actions, either because the inequalities do not hold or there is no possibility of giving money to the opponent. Conversely, when the agent has low prosocial concerns, in some cases, the agent would choose to seize all available resources and then cease accepting additional funds. For instance, if $5/12 \leq \gamma \leq 0.5$, the agent would retain 650 points, leaving the other agent

with 350, instead of taking the entire 900.

Notice, however, that such an inequality aversion model can only accommodate two types of decisions. Hence, we examine the format for a utility function with a continuous structure as described below.

Quadratic inequality aversion

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

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

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

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

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

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

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

We can systematically examine the inequality across all scenarios in our experiment to determine the critical value of ζ at which the agent ceases to accept additional funds under each condition. By solving this inequality for every conceivable situation[8], the resulting solutions yield the values of ζ that satisfy the condition. It is worth mentioning that if the agent commences in a disadvantaged position, there exists no solution with a

[8]It is important to note that our analysis focuses on twin cases; however, an analogous argument can be extended to all cases.

positive ζ.

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

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

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

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

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

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

Balance table

The Table 13 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 13: Balance Table - Average time spent on the experiment, average age, gender, and ethnicity for the groups who face a fine and those who face a fee, with differences and the t-test presented in the last column.

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

7 Order Effects

Table 14 presents an analysis of the amount taken by condition, comparing the order of the session. The following regression model is utilized:

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

Here, *Session* is a dummy variable equal to 1 if the fee is applied in that specific session, *Order* is a dummy variable equal to 1 if the session starts with the treatment condition. Additionally, there is an interaction term evaluating whether the order effect may differ 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 14: Regression (Control) examines order effects for the control conditions, assessing differences in decisions for conditions presented in different orders. Regression (Treatments) investigates order effects for the treatment conditions under various orders.

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

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

Cases & Inequality

Cases

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

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

Here, *Total* indicates the sum of the endowment with the amount taken, and one dummy is used for each case. The results can be observed in Table 15:

	(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 15: 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 extend this analysis, we run the following regression:

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

Here, we analyze the total taken, considering a linear relation for the endowment, and add a dummy to control if the case is dividing 1000 points or 900 points. The results can be observed in Table 16:

	(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 16: 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.

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 17:

	(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 17: 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 18:

	(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 18: 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.

Instructions

Introduction, instructions, and example of comprehension check:

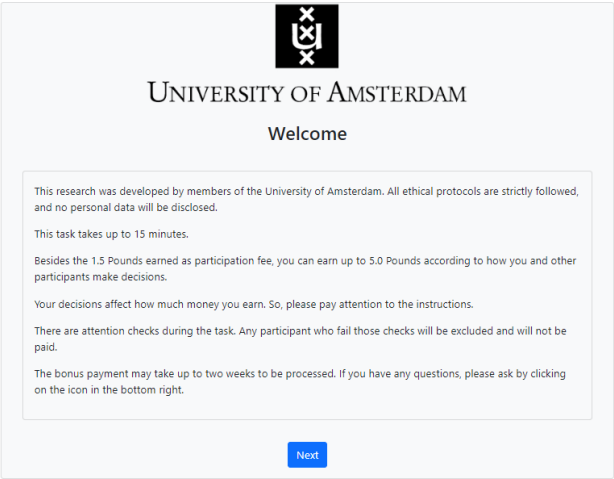

Figure 6: Introduction

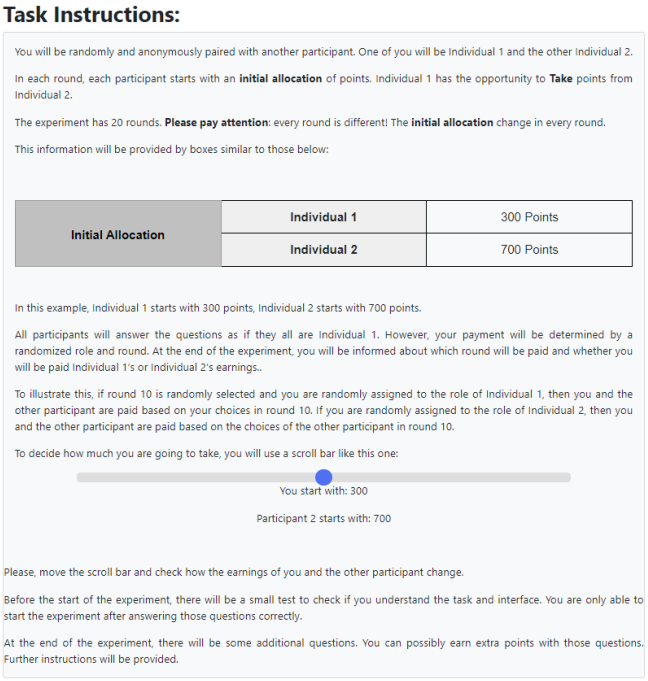

Figure 7: Instructions

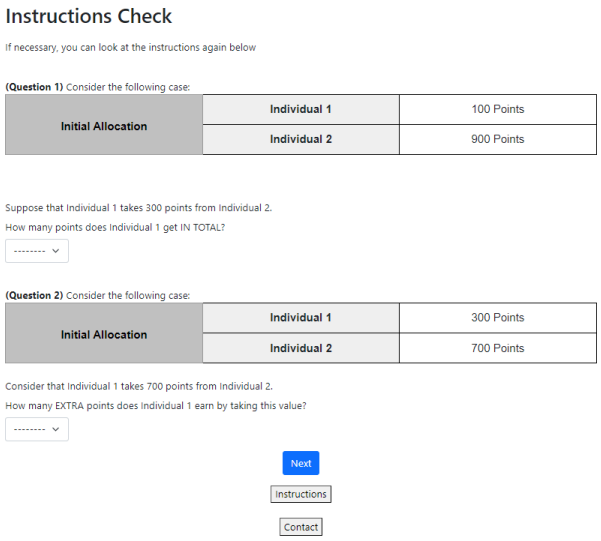

Figure 8: Example - Comprehension check

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

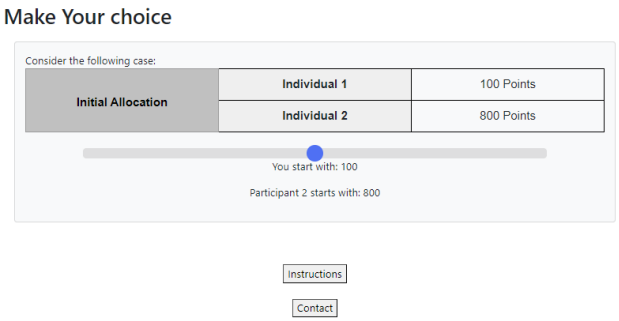

Figure 9: Example: Control Condition

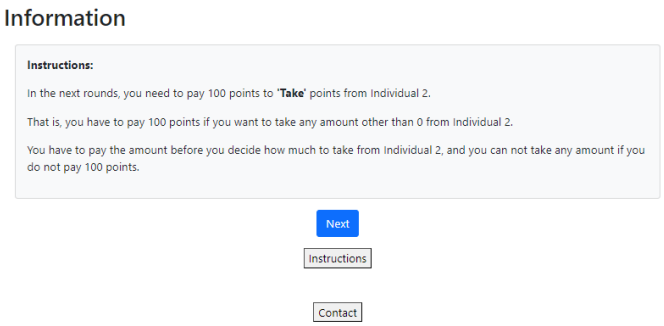

Figure 10: Information - Fine

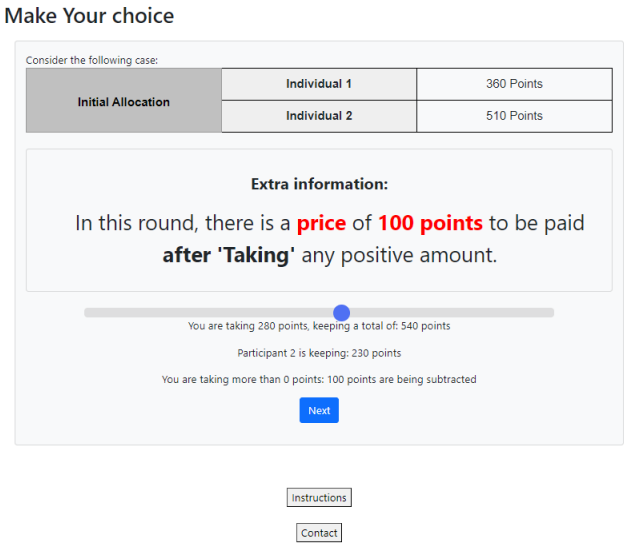

Figure 11: Example: Fine Condition

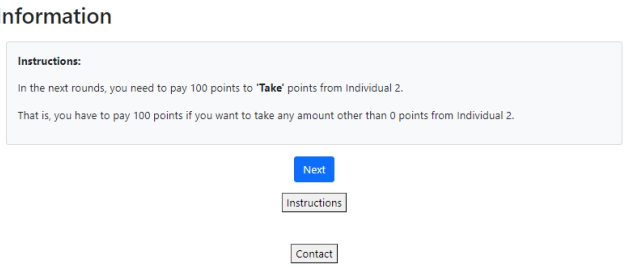

Figure 12: Information - Fee

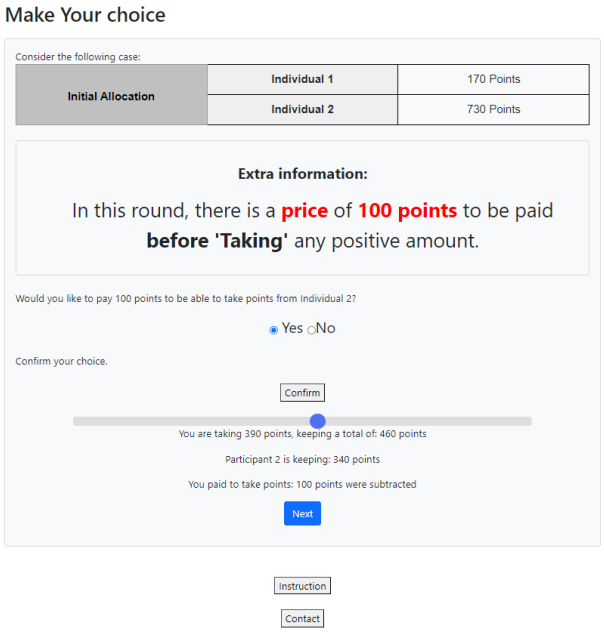

Figure 13: Example: Fee Condition

Social Norms and Entitlement:

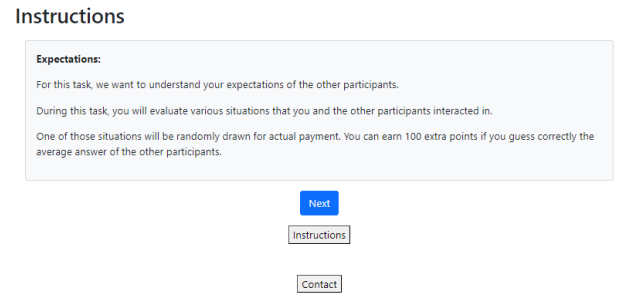

Figure 14: Information - Empirical Expectation

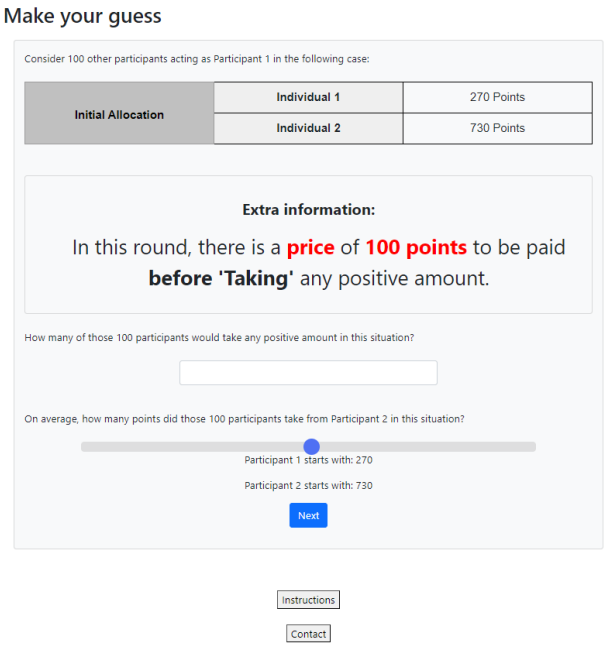

Figure 15: Example: Empirical Expectation

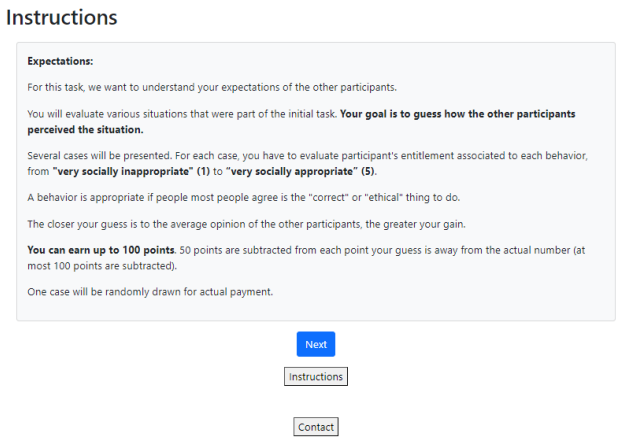

Figure 16: Information - Normative Expectation

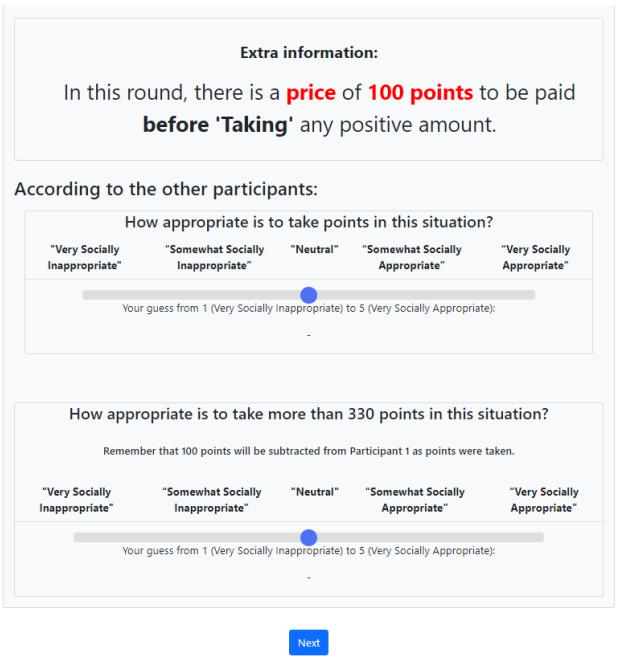

Figure 17: Example: Norm Expectation

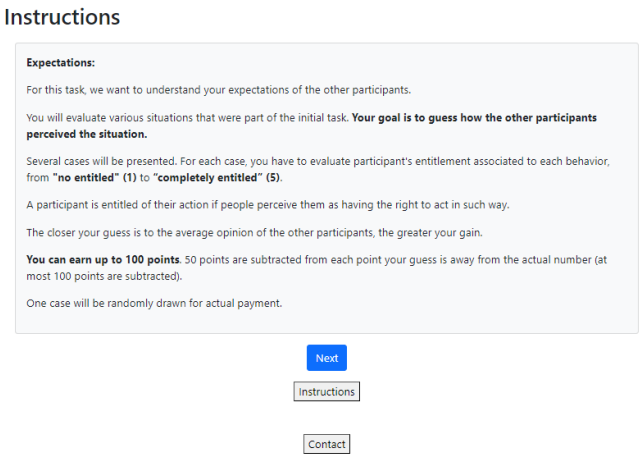

Figure 18: Information - Entitlement

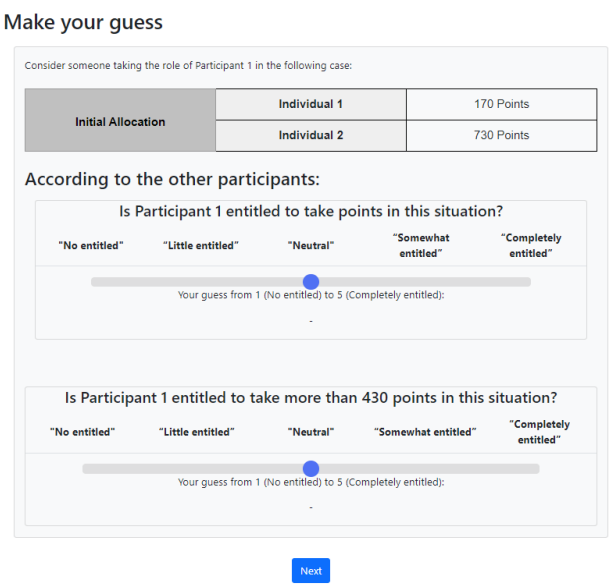

Figure 19: Example: Entitlement