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

Rafael Teixeira

January 30, 2024

JOB MARKET PAPER

Click Here For the 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 formats: one mimicking a 'fine,' imposed *after* taking money, and another mimicking a 'fee,' paid *before* taking money. Our findings reveal that the fee is more effective than the fine in reducing the amount of money taken. In comparison to a situation with no penalty, the fee significantly reduces the aggregate amount taken, whereas the fine shows no significant overall impact. We demonstrate that the differences across conditions can be explained by the individual heterogeneous impact of the penalties: some individuals increase the amount they take when facing a penalty, indicating a crowding-out effect, while others stop taking money when confronted with the penalty, evidence of a crowding-in effect. The fee proves to be more effective in promoting crowding-in than the fine, while crowding-out effects are similar across formats, leading to the overall result. Additionally, we show that the implementation of monetary penalties induces changes in perceived social norms. As individuals conform to these norms, these changes partially explain the crowding-out and crowding-in effects, but they cannot account for the differences across fee and fine.

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 Alexis Belianin, Alejandro Hirmas, Chris Starmer, Fabian Bopp, George Lowenstein, Jan-Willem Stoelhorst, Ro'i Zultan, Sander Onderstal, Shaul Shalvi, Silvia Sonderegger, Simon Gächter, Wendelin Schnedler, Alexander Vostroknutov, 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

Monetary penalties come in various formats and contexts, yet little attention has been paid to how the format of a penalty might impact its efficiency. Our study aims to analyze this matter by examining the behavioral impacts of the format of a monetary penalty, comparing penalties of the same value implemented in two different ways, mimicking features of fees and fines in an experimental setting. There are numerous examples of how penalties are implemented differently. For instance, in environmental legislation, governments often issue emission permits or impose fees on companies, allowing them to emit a specified amount of greenhouse gases before facing penalties. Conversely, companies that violate environmental regulations are frequently subject to fines as a punitive measure imposed after the infraction. Understanding these potential differences is crucial for deriving insights to implement more effective interventions.

Traditional economic theory posits that penalties influence trade-offs by increasing the relative cost of undesirable behavior, thereby reducing its prevalence, and it does not differentiate the potential impacts of penalty formats. However, potential differences might arise as the impact of penalties could extend beyond mere cost-benefit analysis; penalties might also influence prosocial preferences (e.g., Frey and Oberholzer-Gee (1997); Frey (2000); Frey and Jegen (2001)). For example, Gneezy and Rustichini (2000a) describes a crowding-out effect, wherein the implementation of a penalty associated with undesirable behavior leads to an increase in its prevalence. Conversely, Kimbrough and Vostroknutov (2016) suggests that people tend to follow rules, implying that a fine (a new rule) could have a positive impact, as individuals simply adhere to it even without the need for monetary costs, a potential crowding-in effect. Consequently, the literature highlights distinct and inconclusive impacts of penalties on prosocial behavior. Meanwhile, if penalties do influence prosocial preferences, then the format of the penalty might lead to distinguished impacts.

We also investigate the potential causes associated with these changes in prosocial behavior and explore the role of social norms. Social norms (e.g., Janssen and Mendys-Kamphorst (2004); Gneezy, Meier, and Rey-Biel (2011)) have been speculated as a potential explanation for crowding-in and crowding-out phenomena, as the penalty might be perceived as a signal that many individuals are acting in specific ways. Such a signal, perceived as a norm, is used to coordinate toward a "bad" equilibrium. We provide a different approach by drawing insights from Xiao and Bicchieri (2010); Krupka and Weber (2013), which indicate that individuals tend to conform to social norms by directly incorporating norms into their preferences. Additionally, Lane, Nosenzo, and Sonderegger (2023) and Kimbrough and Vostroknutov (2016) demonstrate that implementing a law or rule might induce shifts in social norms. We aim to provide a direct test to determine if the implementation of different monetary penalties leads to different shifts in social norms. As people tend to conform to such norms, it results are different behavioral impacts.

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 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. Participants make decisions under two different conditions: a control condition in which no penalty is implemented and one of the treatment conditions in which one of two different monetary penalties 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, where participants pay *after* any money is taken. The fee condition, and participants face a penalty paid *before* taking any money (i.e., participants have to pay before being able to take money).

We eliminate other confounding factors such as risk concerns to focus solely on a simple timing difference across conditions: Fines are paid *after*, while fees are paid *before*. Both penalties provide the same monetary incentive, and the difference is only a framing effect on the perceived moment of the payment. We aim to disentangle the different impacts of implementing a monetary penalty by precisely controlling for potential trade-off effects, thereby exploring its influence on prosocial preferences.

Following the choices in the dictator game, we assess the participants' social norms. We adopt the terminology 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.

To better explore the behavioral differences between fees and fines, we analyze three impacts: at the aggregate level, examining the average amount of money taken by all subjects; at the extensive margin, referring to the number of instances in which money is taken; and at the intensive margin, considering the amount of money taken, conditional on taking money. Following this analysis, we then compare these behavioral changes with shifts in social norms.

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 condition leads to no significant impact on the amount taken compared to the control, suggesting that this penalty was not effective. In contrast, the fee condition 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.

At the extensive margin, both the fee and fine conditions result in a reduction in the number of instances where money is taken compared to their respective control conditions. The fee condition leads to a significantly greater reduction than the fine. Given that the trade-offs are identical and the subject pool is similar, this difference indicates that the fee condition promotes more prosocial behaviors than the fine condition, suggesting a stronger crowding-in effect.

Hence, we observe that the fee condition is more effective than the fine condition, and this difference reflects the heterogeneous impact that the different penalty formats have on behavior. The crowding-out effects are consistent across conditions, whereas the fee condition leads to higher levels of crowding-in than the fine condition, resulting in a better overall outcome.

The implementation of monetary penalties also induces changes in social norms. 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."

When confronting the behavioral changes alongside shifts in social norms using a mediation model, we observe that social norms can partially explain the treatment effects at both the intensive and extensive margins, thereby partially accounting for the crowding-out and crowding-in behaviors. However, we find no evidence that changes in social norms explain the differences between the fee and fine conditions.

The paper is structured as follows: Section 2 is the theoretical analysis and hypotheses, and Section 3 presents the experimental design. Section 4 contains the results, Section 5 discusses the implications of the findings, and Section 6 concludes.

2 Theory and Hypotheses

This section is divided into three parts: The first part explains traditional economic theories concerning differences in fees and fines within our setting. The second part explores potential behavioral changes, emphasizing trade-offs and potential impacts on prosocial concerns, while also discussing the differing effects fees and fines may have. The third part analyzes the channels and investigates social norms as potential mechanisms for influencing behavioral changes. All hypotheses, the experimental design, and regressions were pre-registered.[1]

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

2.1 The format of a penalty: 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 of engaging in undesirable behavior, they can potentially reduce such behavior (Becker (1968)).

Following this perspective, monetary penalties are implemented in various formats and contexts. For example, 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 of 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)).

Our objective is to investigate whether the format of a monetary penalty influences behavior and to analyze potential mechanisms underlying any observed differences. Therefore, we aim to create a situation in which there is an undesirable behavior to target with the penalties and trying to maintain consistent trade-offs across the different penalties' format.

In our experimental setup, we adapted the traditional dictator game into a "taking game". In an classic dictator game, a participant (dictator) receives some money and has the option to give a share of this money to another participant. In our setting, both participants starts with some money, and one participant has the option to take money from another participant. The original dictator game typically encourages giving behavior, which is generally regarded positively. By reconceptualizing the game in terms of taking, we aimed to simulate a scenario where such behavior is associated with concepts like "stealin" or "greediness."[2]

We implement the monetary penalty associated with the action of taking money to curb this behavior. The penalties are implemented in two different formats: the fine condition and the fee condition. Fines are paid *after* the agent takes any money, while fees are paid *before*, to enable the agent to take any money.

To maintain trade-offs across penalties and focus on the format, we eliminate other confounding factors such as risk concerns and focus on one essential distinction between fees and fines. Hence, both treatment conditions represent the same fixed cost associated with the same behavior, and both fee and fine conditions lead to the same set of potential outcomes. The only difference is the perceived time of payment, *before* or *after*. Notice that in an experimental

[2] It is worth noting that participants indeed perceive taking money as less socially acceptable, particularly in the context of a penalty associated with such behavior. This provides validation for this experimental manipulation.

setting, the actual payment only happens at the end of the experiment, and the format only changes the moment of decision-making, hence, time preferences are not relevant in this context.

A general benchmark for our setting is provided by Gneezy and Rustichini (2000a), where a penalty was imposed on parents picking up their children late from daycare. Similar to this scenario, our setting does not involve a risk component associated with the penalty. Broadly speaking, our experiment captures key aspects of Gneezy and Rustichini (2000a) and explores the effects of paying the penalty before or after the late occurrence. We also simplify the decision-making process to an individual level, which facilitates the examination of behavioral mechanisms, as will be discussed in the following sections.

We can try to analyze the general impacts of the monetary penalties in dictator games in theoretical terms. 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:

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

With the introduction of either a fee or fine, p, as a fixed cost, the agent faces the following problem:

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

Such models cannot differentiate the change across fee and fine, and given that the penalties entail the same trade-offs, classic economic theory would predict that they would yield identical outcomes. In our experimental setup, we impose a fixed before -akin to a fee- and after -akin to a fine- the occurrence of the target behavior. With no further distinctions between them, both penalty types are expected to produce equivalent results.

2.2 Shaping prosocial behavior

Incentives generally affect the trade-offs of 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.

There is a general challenge in disentangling the impacts of trade-offs and changes in prosocial behavior when implementing a penalty. The penalty itself creates an income shock, as individuals must pay for it (in our case, a fixed cost), which should be taken into account in our experimental design.

In our setting, participants engage in multiple rounds of a modified dictator game where money can be taken from the opponent. Across the rounds, participants encounter various cases with different initial endowments with and without the monetary penalty, some these cases represent what we refer to as twin cases.

A twin case consists of options where there is a gap in the initial endowment for the participant who might pay the penalty, and the size of this gap is equivalent to the size of the penalty itself. For example, in case 1, the dictator starts with 'x' points, and the receiver starts with 'y' points. In case 2, its twin case, the dictator starts with '$x+p$' points and the receiver starts with 'y' points, where 'p' is the size of the monetary penalty. This means that if the participant takes money in case 2 and pays the penalty, the set of potential allocations remains the same, allowing us to control for any trade-off differences. Moreover, the participant should take the same amount in those two situations. Consider the notation (x, y) representing the initial endowments of two participants.

Definition 1 - Twin Cases: Given the implementation of a penalty, twin cases are such that (x, y) and $(\hat{x}, y)$, where $\hat{x} = x + p$.

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:

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

The agent, facing a penalty, has three options based on different ζ levels and different initial endowments (as not taking any money means keeping the initial endowment) [3]:

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 case without the penalty.

The third potential behavioral change highlights a crucial aspect of our experiment. For example, consider a dictator starting with 200 points and a receiver with 800 points. In the control condition, with no penalty, someone with strong inequality aversion takes 300 points, resulting in a 500/500 split. By introducing a 100-point penalty, the same agent would take 350 points, resulting in a 450/450 split and taking 50 points more than in the same situation without the penalty. This outcome could be considered a 'more selfish' choice and might be naively interpreted as a change in prosocial concerns, indicative of a crowding-out effect, even without an actual change in the prosocial concerns.

The twin cases control for this income effect/efficiency loss. In the initial endowment 200/800 case in the treatment condition, after the penalty, it can be considered a 100/800 endowment. In the case where the penalty is 100 points, the 100/800 scenario would be its twin case, associated

[3] In Appendix 12, we show the threshold for each case in our treatment.

with the same values given that the agent pays the monetary penalty, and should yield the same decisions: 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, $U(x, y-t)$. 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:

Proposition 1: For twin cases (x, y) and $(\hat{x}, y)$, if a penalty p is implemented and $t^* > 0$, 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. We formulate our base hypothesis based on the twin cases, which leads to clear and precise predictions.

As described, 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:

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

Following the previous discussion, we can also point that there is no significant difference across fee and fine conditions:

Hypothesis 1.1 - Fee vs. Fine: There will be no significant difference in the average amount taken by participants between the fee and fine conditions.

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

Hypothesis 2.1 - Fee vs. Fine: There will be no significant difference in the proportion of cases in which participants take points between the fee and fine conditions.

Hypothesis 3 - Intensive Margin: In the twin cases, if a participant takes points after the introduction of the penalty, there is no difference in the amount taken with or without the penalty.

Hypothesis 3.1 - Fee vs. Fine: In the twin cases, if a participant takes points after the introduction of the penalty, there will be no significant difference in average amount taken by participants between the fee and fine conditions.

As previously described, when considering the trade-offs and the monetary loss due to the penalty, some agents may choose to cease taking money due to a fixed cost. However, as observed by 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, leading to changes at the extensive margin.

For instance, the penalty could be perceived as a form of permission to act, reducing the moral concerns of 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.

Hypothesis 3 directly illustrates proposition 1, and given the twin cases, the choices should be the same across conditions. 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.

2.3 Shaping social norms

Social norms have been described as a key component of these 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 et al. (2011)). These models vary in aspects related to signaling to others, coordination devices, self-image concerns, or even moral considerations. By implementing the game in a dictator setting, we mitigate the influence of strategic interaction and incomplete information. By doing so, we minimize the role of coordination devices and signaling, focusing on one specific way that social norms can affect behavior: conformity.

The introduction of new incentives might trigger different social norms, similar to what is illustrated by Lane et al. (2023). Meanwhile, there is 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. Meanwhile, Ellingsen, Johannesson, Mollerstrom, and Munkhammar (2012); Chang, Chen, and Krupka (2019) shows that the framing of the game might influence the perceived norms associated with the behavior.

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. Hence, we explore this possibility.

We also try to capture a feeling of perceived entitlement. Entitlement may be a factor explaining crowding-out effects (Bénabou and Tirole (2006), Gneezy et al. (2011)). Entitlement is a perceived feeling that an individual holds, often manifesting as the expectation of special treatment, privileges, or rights. 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 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 is 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 propositions:

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

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

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.

3 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.

Participants engage in dictator games, where one 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 an potential 'undesirable behavior.'

We employed the strategic method, with all participants assuming the role of the Dictator. They were informed that they would be randomly matched with another participant and, at the experiment's conclusion, would learn which role they had 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 experiment's end, and participants did not directly interact in any other way besides the amount chosen.

During the experiment, participants played a series of 20 dictator games divided into two blocks: 10 dictator games in the control condition and the same 10 dictator games in one of two treatment conditions:

In the *control* condition, participants could take points from the other participants without any further consequences. In the *treatment* conditions, participants were informed that there was a 100-point penalty associated with taking any money. Across the treatment conditions, we implement the monetary penalty in two different ways, and the specifics of these ways will be provided shortly. Therefore, the impact of each monetary penalty was observed within subjects, while differences in the format of the monetary penalties were observed between subjects.

We varied 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, to investigate if the treatment order might affect behavior.

The 10 dictator games encompassed cases with a range of initial endowments, including scenarios where the dictator began with more money than the receivers and instances where the dictator started with less money than the receiver. Some cases featured the dictator starting with a higher endowment than the receiver, while in others, the receiver started with more endowment. We introduced this variety to check the robustness of any behavioral changes across initial inequality. The order of the different dictator games was randomly presented to the participants.

The endowments aimed to generate twins and enhance decision robustness. Participants consistently allocated either 900 or 1000 points, maintaining consistency across potential choices and contributing to behavioral change robustness. As previously described, twin cases represent dictator games where there is a 100-point gap in the initial endowment of the dictator, which is used to control for income effects. We also included two decoy cases 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 as there is no twin. 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

Notes: 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.

Table 1: Cases (initial endowment for the dictator game)

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 same 100 points penalty in two different ways:

The fee condition captures features associated with a fine and 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.

The fee condition captures features associated with a fee, 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 there is a penalty associated with taking any money for the next decisions. 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**
Before (*'Fee'*)	In this round, there is a **price** of **100 points** to be paid **before 'taking'** any positive amount.
After (*'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 for each treatment condition.

We 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 of 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.

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 the appropriateness of behavior, we adapt it to measure the social perception associated with 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.

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

4 Results

The study involved 201 participants, split between fee and fine conditions, with 101 and 100 participants contributing to a total of 4020 decisions. Our primary focus narrows down to twin cases to address income effects, totaling 1608 observations. Participants also provided information on social norms and perceived entitlement for two twin cases: one where the dictator is behind (Tins 2) and another where the dictator is ahead (Twins 3), resulting in 804 observations for each case.

To ensure robustness, we assessed order effects given the variable session start conditions (control or treatment). No significant differences were observed across the order[4]. Consequently, all corresponding treatment sessions were consolidated for data analysis. The results presented are also robust to other specifications and models, such as the use of hurdle models.

The study's findings are presented in two sections. Section 4.1 delves into the impact of monetary penalties on taking behavior, examining overall changes and breaking them down into extensive and intensive margins. This section also explores behavioral distinctions between fines and fees. In Section 4.2, the study investigates the influence of social norms and entitlement on the amounts taken by participants, analyzing 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. To illustrate any potential behavioral shifts, we can examine the amount of money taken in each condition and each case, as observed in Table 3

[4] Detailed in Appendix 14

Twin	Case	Fine Control Amount Taken	Fine Treatment Amount Taken	Fine Diff	Fee Control Amount Taken	Fee Treatment Amount Taken	Fee 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]

Notes: The average amount taken in each case given and their respective conditions (Fee and Fine) and treatments (control and treatment), along with their differences (p-value for reference). The last column describes the differences-in-differences across fee and fine treatment effects. 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 3: Average amount taken by each case and condition

As the initial endowment increases, the available amount to be taken decreases. Consequently, individuals tend to take lower amounts overall but maintain a higher total share[5].

In the fine condition, when the agent starts with less 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 more 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 than the other participant.

To consolidate the analysis of these behavioral changes, we conducted the following regression model[6].

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

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 of the fine condition and the control condition of the fee condition. [7]

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

Table 4 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.

[5] Further details are provided in the Appendix 15.

[6] For a detailed examination of the impact of the specific cases, check Table 18 in the appendix.

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

	(1 - All data) Take	(2 - Twin cases) 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

Notes: Amount taken (*Take*) regressed on a dummy for *Fee* and *Fine Conditions*. *ControlFine* represents the differences across control conditions associated with fee or fine. Regression (1) uses all observations, while regression (2) uses only the twin cases, controlling for income effects. Random effects at the individual level. Standard errors clustered at the individual level in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$

Table 4: Aggregate treatment effects on the amount taken

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, there is a statistically significant decrease in the amount taken in the fee condition (-27), supporting 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 ($\chi^2(1) = 2.89, p = 0.0894$), indicating that the fee leads to a slightly larger impact than the fine. To illustrate this difference, refer to Figure 1:

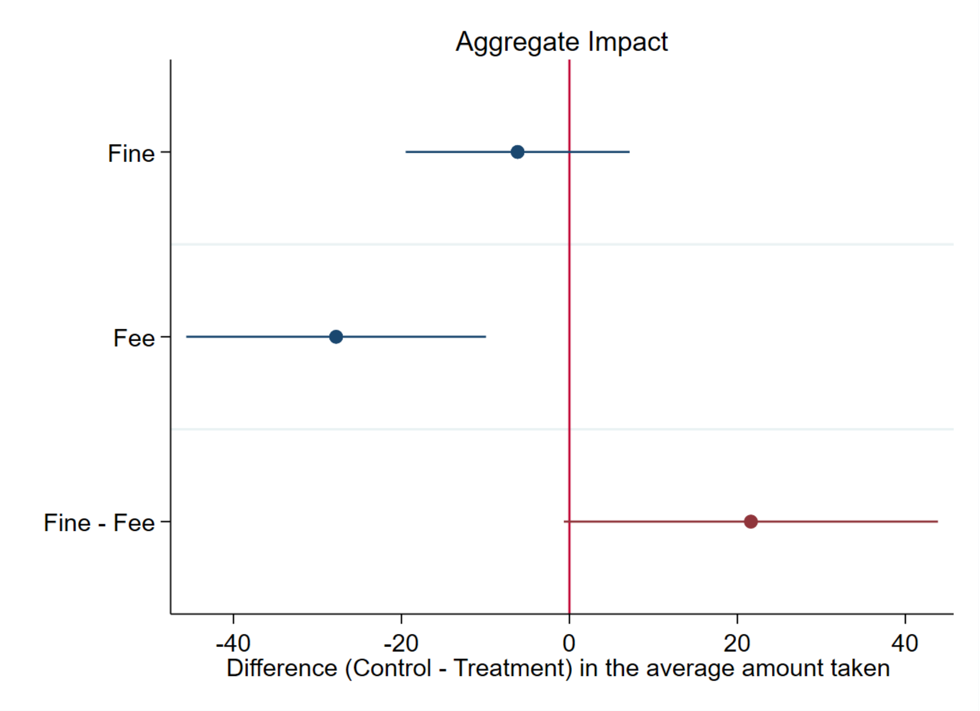


Figure 1: Treatment effects and their 95% confidence intervals on the amount taken at the aggregate level for the twin cases by condition (fee and fine) and their differences.

Result 1 - Aggregate level: *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 1.1 - Aggregate level (Fee vs. Fine): *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 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 check the instances in which any positive amount of money is taken. To do so, we create a dummy variable *Participantion*, equal to 1 if any money is taken in that specific decision. To illustrate any potential behavioral shifts, we can the share of cases in which money is taken for each case, as observed in Table 5:

		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]

Notes: Average number of instances in which money has been taken in each case given and their respective conditions and treatments, along with their differences. *Participantion* is a dummy variable equal to 1 if any money was taken in that decision. The last column describes the differences-in-differences across fee and fine treatment effects. 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 5: Participation by each case and condition

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

In cases where the agent starts with more money than their opponent, a significant portion of participants refrain 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 significantly more drastic impact, leading to even larger reductions that surpass the effect of the fine significantly.

To formally test the behavioral changes, 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 6 presents the results, with Regression (3) using the entire dataset, and Regression (4) focusing on the twin cases.

	(3 - All data)	(4 - Twin cases)
	Participation	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

Notes: The share of instances in which money was taken (*Participation*) is represented as a dummy variable equal to 1 if any money is taken in that decision and regressed on dummy variables for the *Fee* and *Fine Conditions* using a logit model. *ControlFine* represents the differences across control conditions associated with fee or fine. Regression (3) uses all observations, while regression (3) uses only the twin cases, controlling for income effects. Random effects at the individual level. Standard errors clustered at the individual level in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$

Table 6: Extensive margin - instances in which money was taken

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.

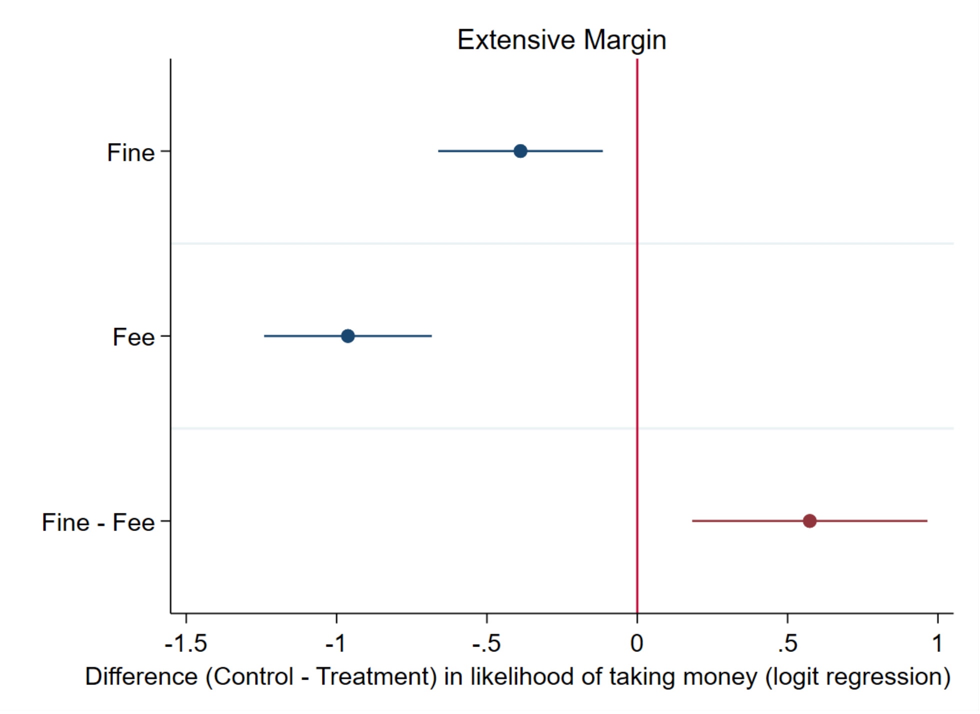


Figure 2: Treatment effects and their 95% confidence intervals on the likelihood of taking money for the twin cases at the aggregate level by condition (fee and fine) and their differences.

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, as it shows that similar agents act more prosocially in the fee condition than in the fine condition.

Result 2 - Extensive Margin: *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 2.1 - Extensive Margin (Fee vs. Fine) : *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.*

Intensive margin:

We proceed with the intensive margin analysis and analyze the amount taken by individuals in the control and treatment condition, conditional on taking any money in the treatment condition, hence, pinpointing the same individual in the same situation. To illustrate these behavioral changes, we can check Table 7:

		Fine			Fee			
Twin	Case	Control Amount Taken	Treatment Amount Taken	Diff	Control Amount Taken	Treatment Amount Taken	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]

Notes: The average amount taken conditional on money being taken (intensive margin) in each case given and their respective conditions (Fee and Fine) and treatments (control and treatment), along with their differences (p-value for reference). The last column describes the differences-in-differences across fee and fine treatment effects. 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 7: Amount taken by each case and condition when money was taken in the treatment condition

We formally assess changes at the intensive margin using the same regression as before, examining variations in the amount taken for each condition, and also focusing on the subsample of participants who continue to take money after the penalty is implemented. While the intensive margin generally concentrates on participants who took any money, as indicated by regression (5), it is crucial to recognize potential differences among participants who took money in the treatment and control conditions, which may introduce an endogenous effect due to varying individuals in each condition.

To address this concern, we specifically chose cases where money was taken in the treatment condition and matched those cases with the corresponding instances 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 pair with the same case for the same individual conditional that individual took money in the treatment condition. Regression (7) pairs with its 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 8 offers additional details.

	(5 - All data)	(6 - 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

Notes: Amount taken (*Take*) conditional on money being taken (intensive margin) regressed on a dummy for *Fee* and *Fine Conditions*. *ControlFine* represents the differences across control conditions associated with fee or fine. Regression (5) uses all observations that money is taken. Regression (6) pairs the cases (control and treatment) for the same participant conditional that participant taking money being taken in the treatment condition. Regression (7) does the same but also pairs the case and its twin case, controlling for income effects. Random effects at the individual level. Standard errors clustered at the individual level in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$

Table 8: Intensive margin

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 3:

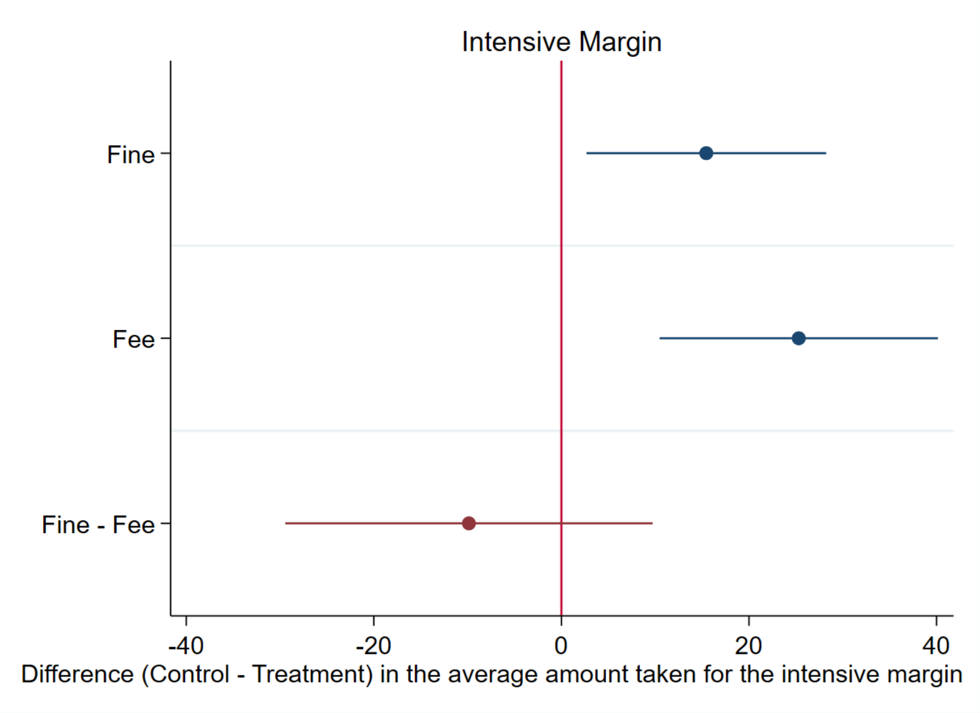


Figure 3: Treatment effects and their 95% confidence intervals on the amount taken for the twin cases conditional on money being taken in the treatment condition (intensive margin) by condition (fee and fine) and their differences.

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 - Intensive Margin: *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 3.1 - Intensive Margin (Fee vs. Fine): *At the intensive margin, there are no significant differences between the treatment effects of* ***fee*** *and* ***fine****.*

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 significant 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.[8] 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.

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 (social norms); and perceived entitlement.

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

The first aspect reflects the extensive margin: To analyze the potential changes, we employ the same regression as in the previously, bu adjusting the dependent variable for each measure of social norm/entitlement, *Empirical*, *Normative* or *Entitlement*.

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 (1.0 to 4.0), and for perceived entitlement, we ask participants about their perception of how entitled others were to take any amount (1.0 to 4.0).

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] The analysis delving into the relationship between inequality and behavioral changes is discussed and illustrated in Appendix D.

	(8) Empirical	(9) Normative	(10) Entitlement
Fine	-5.866***	-0.1812***	-0.1010*
	(1.239)	(0.0526)	(0.0601)
Fee	-4.860***	-0.2010***	-0.1550***
	(1.484)	(0.0477)	(0.0442)
ControlFine	3.476	-0.0919	-0.0693
	(2.432)	(0.0838)	(0.0962)
Constant	65.95***	3.430***	3.250***
	(1.645)	(0.0635)	(0.0721)
N	804	804	804

Notes: Changes in social norms associated with the likelihood or act of taking any amount of money (extensive margin) regressed on a dummy for *Fee* and *Fine Conditions*. *ControlFine* represents the differences across control conditions associated with fee or fine. Regression (8) analyzes the empirical expectation (share of individuals who would take money). Regression (9) analyzes the normative expectations (appropriateness levels of taking any money). Regression (10) checks the perceived entitlement of taking any amount of money. Random effects at the individual level. Standard errors clustered at the individual level in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$

Table 9: Changes in social norms for the extensive margin

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 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.[9]

	(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

Notes: Changes in social norms associated with the how much or act of taking larges amount of money (intensive margin) regressed on a dummy for *Fee* and *Fine Conditions*. *ControlFine* represents the differences across control conditions associated with fee or fine. Regression (11) analyzes the empirical expectation (expected amount that individuals would take). Regression (12) analyzes an weighted empirical expectations (expectations of how much × expectations of how likely is to take). Regression (13) analyzes the normative expectations (appropriateness levels of taking larges money). Regression (14) checks the perceived entitlement of taking largers amount of money. Random effects at the individual level. Standard errors clustered at the individual level in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$

Table 10: Changes in social norms for the intensive margin

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 of 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 increases the perceived entitlement to take larger amounts of money, while it does not affect the fine condition.

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

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 check 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 measured (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:

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

The additional variables, *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.

The impact of social norms might differ from the fee and fine condition, and hence, we use the following regression to control for this aspect:

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

This regression adds an interaction term between the *Fee* dummy that captures the treatment condition, and each social norm. Such interaction terms would differentiate any potential difference impact of each measure on the behavior across the treatment conditions.

These models represent a mediation model, similar to those suggested by Howell (1992) and others. The general idea is that changes in social norms can explain the changes in behavior, and hence the changes are correlated. If this is the case, the coefficients associated with the social norms would partially capture the treatment effects.

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[10]. 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.

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

	(15)	(16)	(17)	(18)	(19)	(20)
	Participation	Participation	Participation	Take	Take	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

Notes: Social norms serve as potential channels for understanding behavioral changes. Regressions 15, 16, and 17 address the extensive margin, where we regress a dummy variable indicating instances of money being taken (*Participation*) on dummies for *Fee* and *Fine* Conditions using a logit model. Regression (15) replicates the previous analysis of the extensive margin conducted in regression (4) using a subsample for which social norms have been measured. In Regression (16), we augment the model by incorporating social norms (Empirical and Normative) and perceived entitlement (*Entitlement*) as explanatory variables. Regression (17) further extends the model by introducing an interaction between the treatment condition dummy (*Fee*) and social norms. Regressions 18, 19, and 20 address the intensive margin, focusing on the likelihood of taking money (*Take*) and regressing it on dummies for *Fee* and *Fine* Conditions using a linear model with random effects. Regression (18) replicates the previous intensive margin analysis conducted in regression (7) using the subsample for which social norms have been measured. In Regression (19), we expand the model to include social norms (Empirical and Normative) and perceived entitlement (*Entitlement*) as explanatory variables. Regression (20) further extends the model by introducing an interaction between the treatment condition dummy (*Fee*) and social norms. *ControlFine* represents the differences across control conditions associated with fee or fine. Regression (3) utilizes all observations, while regression (3) employs only the twin cases, controlling for income effects. Random effects at the individual level. Standard errors are clustered at the individual level in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table 11: Social Norms as potential channels

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, 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. This result indicates that the impact of social norms on behavior is fairly consistent across fee and fine.

Given such results, we analyze the changes in the coefficients for the treatment effects and their differences change across the regressions to check if social norms can be a potential mediator for behavioral change:

When comparing the results of regression (15) and (16) to analyze the extensive margin, the coefficients of 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 of the fine treatment effect are not significantly different ($\chi^2(1) = 1.64, p = 0.2000$). However, this result might partially be attributed to the fact that the coefficient itself was not significant in the replication (regression 17), leaving less room for the influence of social norms. Regarding the Fee condition, the coefficient change is marginally significant ($\chi^2(1) = 3.04, p = 0.0812$).

Again, the difference between fee and fine was not significantly explained by changes in social norms and entitlement ($\chi^2(1) = 0.26, p = 0.6084$). These results indicate that the drop in coefficients for the fee condition is significant, while the decrease for the fine condition is illustrative but not statistically significant. Hence, social norms partially explain the treatment effects, especially for the fee condition.

Result 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 and fine conditions.

Result 5.1: Changes in the 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 account for the differences between the treatment conditions (fee vs. fine).

5 Discussion

We compare the impact of monetary penalties of different formats, mimicking features of fines and fees, on behavior, aiming to identify differences in effectiveness. We also pinpoint the effects of monetary penalties on prosocial concerns, aiming to understand the actual impacts of monetary penalties on behavior and their relation to the context. Lastly, we attempt to analyze the mechanisms behind such changes, examining the role of social norms in these behavioral shifts.

We employed modified dictator games, allowing individuals to take money from others, and introduced fees or fines aiming to reduce the amount of money being taken. We designed our experiment's penalties to avoid other confounding factors, focusing on a timing difference between fees and fines: Fees impose penalties *before* the action, while fines impose *after*. Hence, the penalties reflect the same fixed cost and trade-offs, and should lead to the same impacts.

However, we systematically observed differences between fees and fines, with the fine condition being inefficient in significantly changing the amount taken compared to the situation with no penalty, while the fee condition was effective, resulting in a significant reduction. The differences across conditions were significant, illustrating that the format of the penalty affects its efficiency, suggesting potential changes in legislation targeting the format of monetary penalties to increase efficiency.

The differences across fees and fines reflect their impact on prosocial preferences. There is a myriad of literature describing how monetary penalties might potentially worsen the situation, as described in crowding-out effects (e.g., Frey and Jegen (2001); Frey (2000); Frey and Oberholzer-Gee (1997); Gneezy and Rustichini (2000a); Festré and Garrouste (2015)). Meanwhile, Kimbrough and Vostroknutov (2016, 2018) describe the potential for crowding-in effects, as people conform to rules and increase prosocial behavior even when it is costly to do so in order to comply with a rule. Both behaviors are mutually exclusive, and this has been a puzzle in the literature if and how penalties might work.

Our results answer this puzzle as they indicate that the penalties have heterogeneous impacts on individuals: Some agents exhibit crowding-out effects, increasing the amount of money they take when a penalty is implemented, while others display crowding-in effects, ceasing to take money even when they would benefit from doing so. The overall impact of the penalty represents the balance of these forces, which change across formats, fees, fines, and contexts.

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 (80% to 65%) in the number of cases where money is taken, while the fine treatment results in a 5% reduction (80% to 75%). This difference is significant, indicating that people were acting more prosocially in the fee condition, given the same trade-offs and similar samples, showing a stronger crowding-in effect.

Meanwhile, participants who persist in taking money after the penalty's implementation exhibit a significant increase in the amount they take, 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.

The impact at the aggregate level reflects the combined effects of these heterogeneous impacts. 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.

Hence, our results shed light on the conflict between crowding-out effects (e.g., Frey and Jegen (2001); Frey (2000); Frey and Oberholzer-Gee (1997); Gneezy and Rustichini (2000a); Festré and Garrouste (2015)) and crowding-in effects (e.g., Kimbrough and Vostroknutov (2016, 2018)). When observing individual behaviors, the behaviors are not mutually exclusive, as individuals respond differently, with some agents fitting each theory. Moreover, the format itself affects the size of these effects; the fine conditions, as not effective, could be considered a crowding-out effect as observed at the aggregate level, while the fee condition, by reducing the amount taken, demonstrates a simple direct impact of monetary penalties. However, the aggregate level change reflects the balance of each crowding-in and crowding-out effect.

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. Contemporary approaches, such as carbon markets, might bring about even more moral changes, and might worst outcomes. The format of a market, a penalty and an intervention might lead to different outcomes, similar to what is observed by Falk and Szech (2013); Bartling, Weber, and Yao (2015); Bartling, Fehr, and Özdemir (2023). Our results emphasize the importance of analyzing the moral impacts of each setting to create truly effective interventions.

To delve into the mechanisms driving behavioral changes, we conduct an analysis of the role played by social norms. Social norms (e.g., Janssen and Mendys-Kamphorst (2004), Gneezy et al. (2011)) have been posited as a potential explanation for both crowding-in and crowding-out phenomena, as the implementation of penalties may be perceived as a signal that many individuals are acting in specific ways. This signal might be considered a norm and facilitates coordination toward, for example, a "bad" equilibrium, leading to crowding-out effects.

Similar discussions have been put forth by Ellingsen and Mohlin (2022), Capraro and Perc (2021), Bénabou and Tirole (2006), and Bénabou and Tirole (2003), aiming to integrate self-image concerns or moral duties into preferences to elucidate behavioral changes. We build upon this approach by drawing insights from Xiao and Bicchieri (2010) and Krupka and Weber (2013), which suggest that individuals often conform to social norms by directly incorporating them into their preferences. Meanwhile, Lane et al. (2023) show that implementation of laws can influence social norms.

Hence, we direct test if the implementation of the (different) penalties, lead to (different) shifts in social norms, and lead to (different) behavioral change). If the social norm "improves" due to the penalty, a crowding-in would be observed, if the social norm "deteriorates", a crowding-out would be observed.

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." However, such changes are not significantly differently across fee and fine condition.

The only difference observed across conditions were in our novel measure for entitlement. This result indicate that there are small differences across entitlement and social norms, and entitlement might be changing differently across conditions.

We also observe a positive correlation between norms/entitlement 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.

When directly testing the relation between behavioral changes and social norms/entitlement using a mediation model. Our regression model shows that the treatment effects associated with the fee and fine are partially explained by changes in social norms/entitlement. This suggests that the shifts in social norms can explain crowding-in and crowding-out effects to some extent.

Our results indicate that changes in social norms can partially explain the crowding-in and crowding-out effects. These results provide further context to models such as those by Bénabou and Tirole (2006), Ellingsen and Mohlin (2022), and Capraro and Perc (2021), which use social norms to capture moral perspectives and social image concerns. Moreover, they resonate with results presented by Ellingsen et al. (2012) and Chang et al. (2019), as well as other experiments that demonstrate how the framing of the game affects the expectations associated with that context. Hence, framing effects are an important tool to understand how social image concerns and morals are constructed, and behavioral changes might be driven not only by the context given, whether it's a fee or a fine, but also by the framing, without the need for additional incentives.

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

Our study not only 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 effects on prosocial concerns, leading to unexpected results. Such impacts on prosocial concerns 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 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 is 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, explaining both crowding-in and crowding-out 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.

Bartling, B., Fehr, E., & Özdemir, Y. (2023). Does market interaction erode moral values? *Review Of Economics and Statistics*, *105*(1), 226–235.

Bartling, B., Weber, R. A., & Yao, L. (2015). Do markets erode social responsibility? *The Quarterly Journal of Economics*, *130*(1), 219–266.

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.

Chang, D., Chen, R., & Krupka, E. (2019). Rhetoric matters: A social norms explanation for the anomaly of framing. *Games and Economic Behavior*, *116*, 158–178.

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.*

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.

Festré, A., & Garrouste, P. (2015). Theory and evidence in psychology and economics about motivation crowding out: A possible convergence? *Journal of Economic Surveys*, *29*(2), 339–356.

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

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

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

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

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

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

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

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

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

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

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

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. (2023). Law and norms: Empirical evidence. *American Economic Review*, *113*(5), 1255–1293.

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

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*	$\zeta < \frac{2}{3}$	$900 - 900\zeta$	$\zeta > \frac{5}{12}$

Notes: 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.

Table 12: Thresholds for inequality aversion and the specific behavioral changes in each case.

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[11], 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 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:

[11] Note that our analysis focuses on twin cases; however, an analogous argument can be extended to all 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
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

Average time spent on the experiment, average age, gender, and ethnicity for both groups subjected to fines and those subjected to fees, along with their respective standard deviations in parentheses. The last column illustrates the differences across treatments and describes their p-value in brackets. * $p<0.05$, ** $p<0.01$, *** $p<0.001$. Standard deviations are presented in parentheses, and t statistics are enclosed in brackets.

Table 13: Balance Table

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:

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

Here, *Fee* 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)	(Treatments)
	Take	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

Notes: The Regression (Control) analyzes order effects for the control conditions, assessing differences in decisions when conditions are presented in different orders. Regression (Treatments) investigates order effects for the treatment conditions presented in various orders. The variable *Fee* represents the distinction between fee and fine conditions, while *order* captures differences if the session starts with a treatment or control condition, including the interaction term between *order* and *fee*. Standard errors, clustered at the individual level, are provided in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$

Table 14: Order effects

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, $case_i$ is used for each case, i. 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

Notes: *Total* (Endowment + amount taken) and instances that money is taken (*Participantion*) regressed on dummies for each case, represented by the numbers. 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 for participation. 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)	(2)	(3)
	Total	Total	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: Cases impacting the total amount kept by the participant

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 compare the same cases (twin cases), so this observed tendency does not 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 compare the twin cases, which control for income effects and serve as the main benchmark of our results. To do so, we use the following regressiion:

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

$$\beta_4 Ahead + \beta_5 Ahead \times Fee + \beta_6 Ahead \times Fine + \epsilon_{i,r}$$

We begin by analyzing the aggregate results, which can be observed in Table 18:

	(1 - Behind) Take	(2 - Ahead) Take	(3 - All) Take
ControlFine	-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

Notes: Amount taken (*Take*) regressed on a dummy for *Fee* and *Fine Conditions*. *ControlFine* represents the differences across control conditions associated with fee or fine. *Ahead* is a dummy capturing if the agent starts with more money than their opponent, and the an interaction term between *ahead* and the treatment conditions. 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. Random effects at the individual level. Standard errors clustered at the individual level in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$

Table 17: Aggregate Impact by inequality

The results indicate that the Fee condition is effective only when the agent is in a leading position, showing no significant impact or differences under other circumstances.

Additionally, we analyze the changes at the extensive margin in Table **??**:

	(4 - Behind) Participation	(5 - Ahead) Participation	(6 - All) Participation
ControlDiff	0.783	-0.101	-0.0322
	(1.810)	(1.111)	(0.553)
Fine	-1.959	-1.111**	-1.186
	(1.451)	(0.527)	(1.111)
Fee	-2.296	-3.557***	-2.773**
	(1.473)	(0.769)	(1.205)
Ahead			-6.589***
			(1.116)
Fine × Ahead			0.317
			(1.120)
Fee × Ahead			0.484
			(1.294)
Constant	7.294***	1.495*	7.688***
	(1.830)	(0.875)	(1.216)
N	804	804	1608

Notes: Instances that money is taken (*Participation*) regressed on a dummy for *Fee* and *Fine Conditions*. *ControlFine* represents the differences across control conditions associated with fee or fine. *Ahead* is a dummy capturing if the agent starts with more money than their opponent, and the an interaction term between *ahead* and the treatment conditions. Regression (4) describes the impact of treatment on the amount taken for cases in which the dictator starts behind, Regression (5) for cases in which the dictator starts ahead, and Regression (6) includes all data. Random effects at the individual level. Standard errors clustered at the individual level in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$

Table 18: Extensive margin by inequality

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 19:

	(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

Notes: Amount taken (*Take*) conditional on money being taken in the treatment condition regressed on a dummy for *Fee* and *Fine Conditions*. *ControlFine* represents the differences across control conditions associated with fee or fine. *Ahead* is a dummy capturing if the agent starts with more money than their opponent, and the an interaction term between *ahead* and the treatment conditions. 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. Random effects at the individual level. Standard errors clustered at the individual level in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$

Table 19: Intensive margin by inequality

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.

Hurdle Models

Another method for exploring treatment effects on the intensive and extensive margins involves employing hurdle models. Essentially, these models use a two-staged regression, one for selection (extensive margin) and another using a linear model for the action (intensive margin). Here, we examine the results obtained through such models. The initial regression assesses treatment effects on the intensive and extensive margins using all available data:

	(1) Take
Fine	52.35***
	(12.37)
Fee	99.04***
	(12.84)
ControlFine	-4.064
	(12.28)
Constant	320.8***
	(9.001)
Selection	
Fine	-0.232***
	(0.0616)
Fee	-0.610***
	(0.0606)
ControlFine	-0.0287
	(0.0639)
Constant	0.863***
	(0.0455)
N	4020

Notes: Hurdle model for Participation and Amount Taken. Amount taken (*Take*) regressed on a dummy for *Fee* and *Fine Conditions*. *ControlFine* represents the differences across control conditions associated with fee or fine. Random effects at the individual level. Standard errors clustered at the individual level in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, ***

Table 20: Hurdle Model

The second and third regressions check the impacts of the social norms on the treatment effects:

	(2) Take	(3) Take
Fine	23.64	6.711
	(24.59)	(18.19)
Fee	66.84***	29.73
	(25.47)	(18.91)
ControlFine	-0.988	0.770
	(24.30)	(17.91)
Empirical Intensive		0.608***
		(0.0400)
Normative Intensive		1.877**
		(0.743)
Entitlement Intensive		0.406
		(0.662)
Constant	345.6***	88.04***
	(17.54)	(24.36)
Selection		
Fine	-0.218	-0.131
	(0.140)	(0.174)
Fee	-0.551***	-0.688***
	(0.137)	(0.172)
ControlFine	-0.0116	-0.0615
	(0.145)	(0.181)
Empirical Extensive		0.0174***
		(0.00210)
Normative Extensive		0.0301***
		(0.00710)
Entitlement Extensive		0.0179***
		(0.00626)
N	804	804

Notes: We employ a Hurdle model to analyze Participation and Amount Taken, considering social norms as potential channels. The Amount taken (*Take*) is regressed on dummy variables for *Fee* and *Fine Conditions*. *ControlFine* accounts for differences across control conditions associated with fee or fine. In Regression (2) and (3), we examine treatment effects (fine and fee) on the amount taken, capturing the intensive margin, and treatment effects (fine and fee) for the selection model, capturing the extensive margin using the situations in which social norms were measured. Additionally, in Regression (3), we include social norms to capture their impacts. We utilize random effects at the individual level, with standard errors clustered at the individual level in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table 21: Hurdle Model and Channesl

The observed results align with those described in the main text, indicating that both fees and fines result in reductions at the extensive margins but increases at the intensive margin. Additionally, social norms exhibit a positive correlation with selection and amount taken, partially capturing the treatment effects.

Who are those who ceased with taking money:

We also investigate the behavior of agents who cease taking money during the control condition, i.e., how much they take in the control condition for the twin case in which they stop taking money in the treatment condition. Figure 4 displays the distribution of the amount taken for the same respective control conditions in which the agent did not take money in the treatment condition.

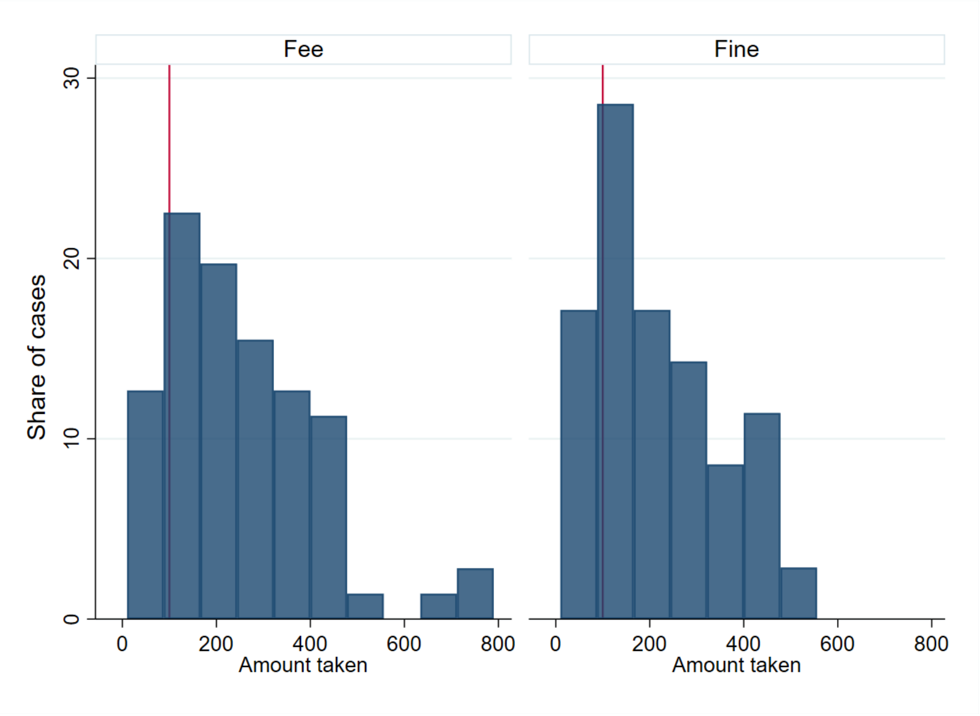


Figure 4: The distribution of the amount taken among those who did not take money in the treatment conditions. On the left side, the amount taken in the control condition by those who did not take money in the fee treatment. On the right side, the same information is presented for the fine treatment.

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 5 shows the distribution of these values.

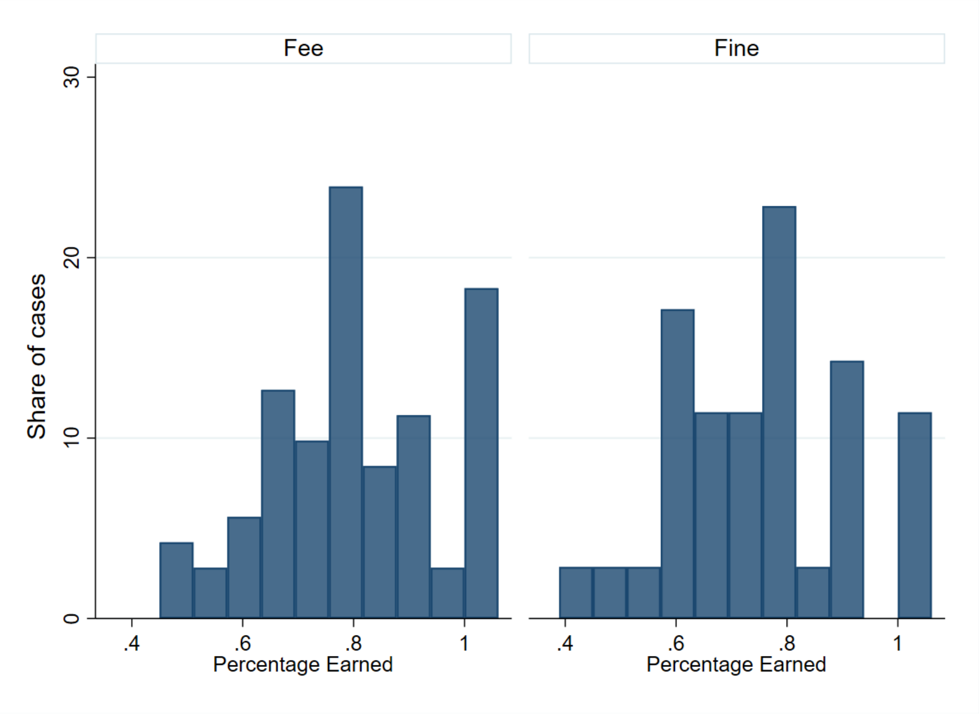
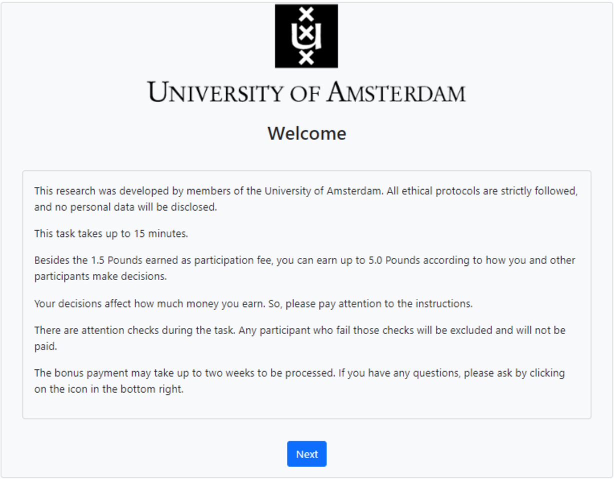


Figure 5: The distribution of the total share kept among those who did not take money in the treatment condition is shown on the left side. On the left side, the share kept in the control condition by those who did not take money in the fee treatment is displayed, while on the right side, the same information 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 of the fee treatment, dictators obtain 100% of the money in 18.3% of cases, while in control conditions of the fine treatment, this occurs in 11.43% of cases, and these individuals decide to stop taking any money after the penalty is imposed. These significant reductions in the amount taken serve as evidence for a crowding-in effect. The agent's drastic reduction in the amount taken indicates that the monetary penalty indeed leads to an increase in prosocial concerns. This is evident as they exhibit little prosocial behavior by taking larger amounts in the control condition, but demonstrate a higher level of prosociality by taking zero in the treatment condition.

Instructions

Introduction, instructions, and example of comprehension check:



UNIVERSITY OF AMSTERDAM

Welcome

This research was developed by members of the University of Amsterdam. All ethical protocols are strictly followed, and no personal data will be disclosed.

This task takes up to 15 minutes.

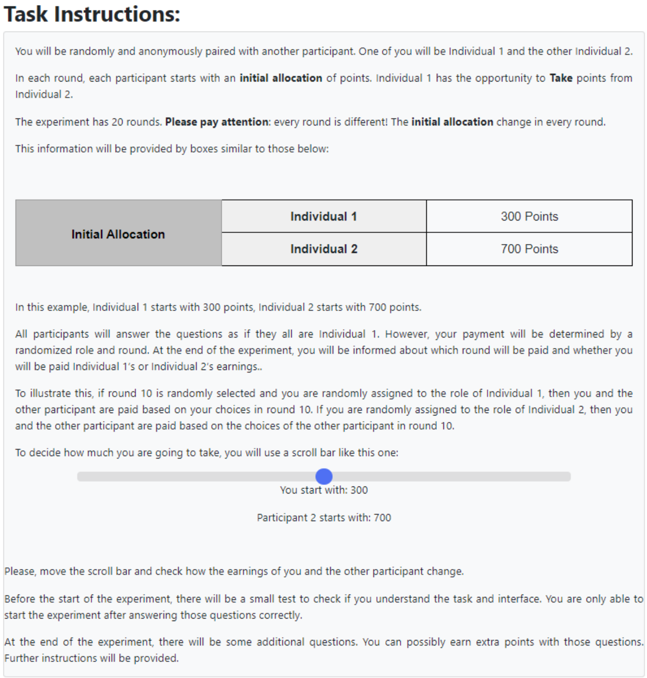
Besides the 1.5 Pounds earned as participation fee, you can earn up to 5.0 Pounds according to how you and other participants make decisions.

Your decisions affect how much money you earn. So, please pay attention to the instructions.

There are attention checks during the task. Any participant who fail those checks will be excluded and will not be paid.

The bonus payment may take up to two weeks to be processed. If you have any questions, please ask by clicking on the icon in the bottom right.

Figure 6: Introduction



Task Instructions:

You will be randomly and anonymously paired with another participant. One of you will be Individual 1 and the other Individual 2.

In each round, each participant starts with an **initial allocation** of points. Individual 1 has the opportunity to **Take** points from Individual 2.

The experiment has 20 rounds. **Please pay attention**: every round is different! The **initial allocation** change in every round.

This information will be provided by boxes similar to those below:

Initial Allocation	**Individual 1**	300 Points
	Individual 2	700 Points

In this example, Individual 1 starts with 300 points, Individual 2 starts with 700 points.

All participants will answer the questions as if they all are Individual 1. However, your payment will be determined by a randomized role and round. At the end of the experiment, you will be informed about which round will be paid and whether you will be paid Individual 1's or Individual 2's earnings..

To illustrate this, if round 10 is randomly selected and you are randomly assigned to the role of Individual 1, then you and the other participant are paid based on your choices in round 10. If you are randomly assigned to the role of Individual 2, then you and the other participant are paid based on the choices of the other participant in round 10.

To decide how much you are going to take, you will use a scroll bar like this one:

You start with: 300

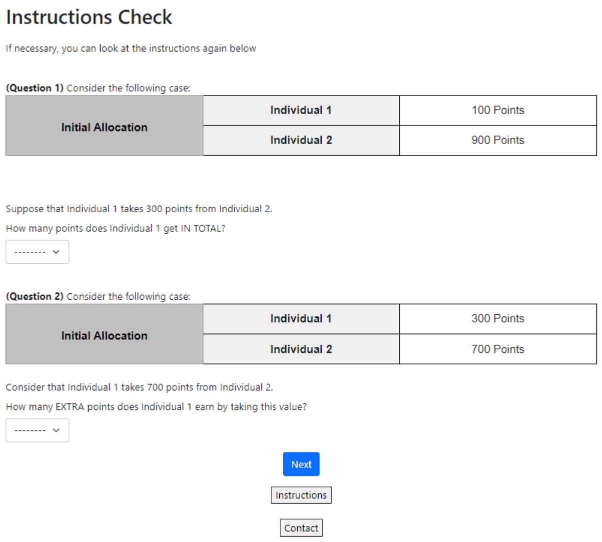
Participant 2 starts with: 700

Please, move the scroll bar and check how the earnings of you and the other participant change.

Before the start of the experiment, there will be a small test to check if you understand the task and interface. You are only able to start the experiment after answering those questions correctly.

At the end of the experiment, there will be some additional questions. You can possibly earn extra points with those questions. Further instructions will be provided.

Figure 7: Instructions



Instructions Check

If necessary, you can look at the instructions again below

(Question 1) Consider the following case:

Initial Allocation	**Individual 1**	100 Points
	Individual 2	900 Points

Suppose that Individual 1 takes 300 points from Individual 2.

How many points does Individual 1 get IN TOTAL?

(Question 2) Consider the following case:

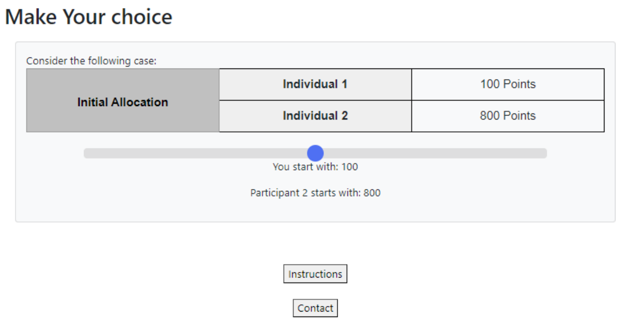
Initial Allocation	**Individual 1**	300 Points
	Individual 2	700 Points

Consider that Individual 1 takes 700 points from Individual 2.

How many EXTRA points does Individual 1 earn by taking this value?

Figure 8: Example - Comprehension check

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



Make Your choice

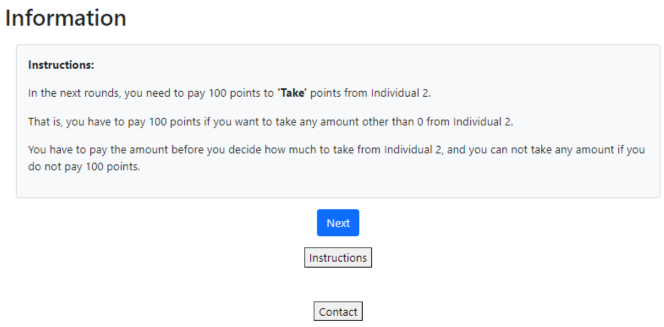
Consider the following case:

Initial Allocation	**Individual 1**	100 Points
	Individual 2	800 Points

You start with: 100

Participant 2 starts with: 800

Figure 9: Example: Control Condition



Information

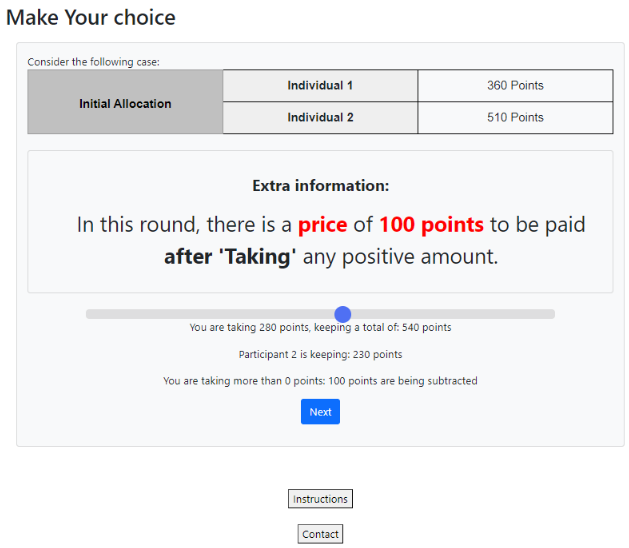
Instructions:

In the next rounds, you need to pay 100 points to '**Take**' points from Individual 2.

That is, you have to pay 100 points if you want to take any amount other than 0 from Individual 2.

You have to pay the amount before you decide how much to take from Individual 2, and you can not take any amount if you do not pay 100 points.

Figure 10: Information - Fine



Make Your choice

Consider the following case:

Initial Allocation	**Individual 1**	360 Points
	Individual 2	510 Points

Extra information:

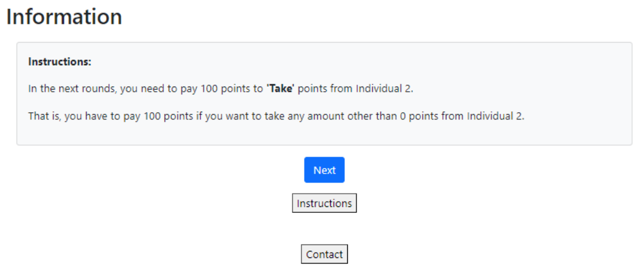
In this round, there is a **price** of **100 points** to be paid **after 'Taking'** any positive amount.

You are taking 280 points, keeping a total of: 540 points

Participant 2 is keeping: 230 points

You are taking more than 0 points: 100 points are being subtracted

Figure 11: Example: Fine Condition



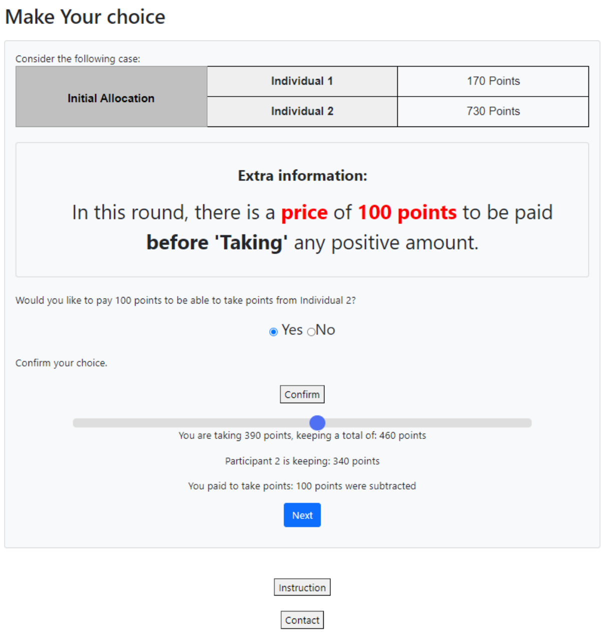
Information

Instructions:

In the next rounds, you need to pay 100 points to '**Take**' points from Individual 2.

That is, you have to pay 100 points if you want to take any amount other than 0 points from Individual 2.

Figure 12: Information - Fee



Make Your choice

Consider the following case:

Initial Allocation	**Individual 1**	170 Points
	Individual 2	730 Points

Extra information:

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

Would you like to pay 100 points to be able to take points from Individual 2?

● Yes ○ No

Confirm your choice.

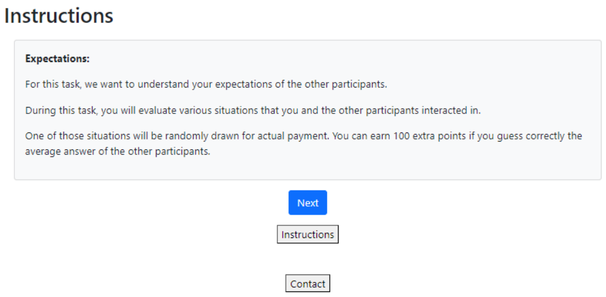
You are taking 390 points, keeping a total of: 460 points

Participant 2 is keeping: 340 points

You paid to take points: 100 points were subtracted

Figure 13: Example: Fee Condition

Social Norms and Entitlement:



Instructions

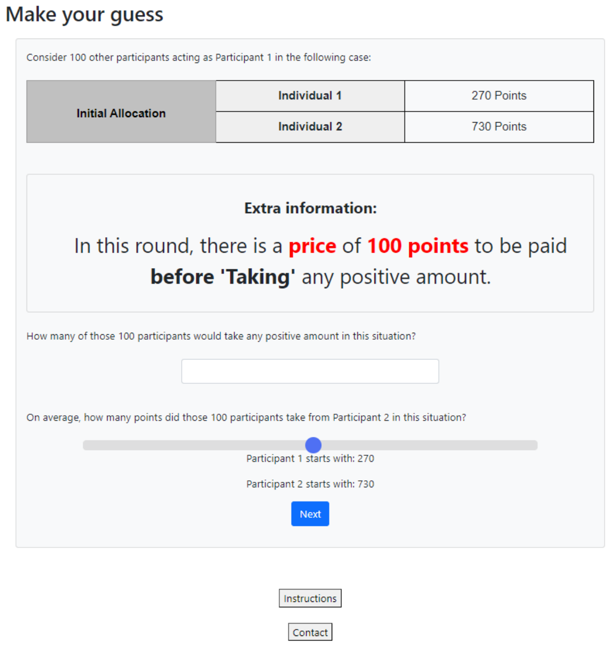
Expectations:

For this task, we want to understand your expectations of the other participants.

During this task, you will evaluate various situations that you and the other participants interacted in.

One of those situations will be randomly drawn for actual payment. You can earn 100 extra points if you guess correctly the average answer of the other participants.

Figure 14: Information - Empirical Expectation



Make your guess

Consider 100 other participants acting as Participant 1 in the following case:

Initial Allocation	**Individual 1**	270 Points
	Individual 2	730 Points

Extra information:

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

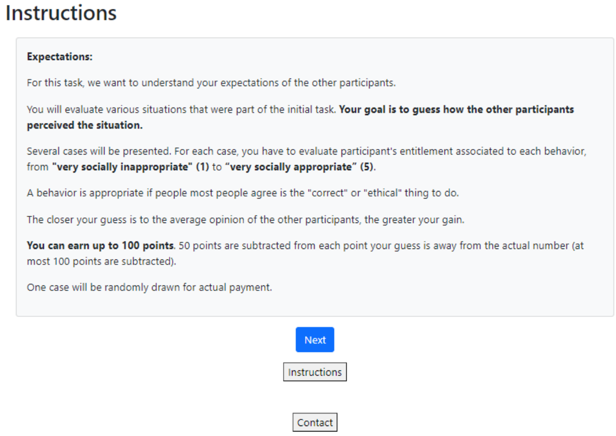
How many of those 100 participants would take any positive amount in this situation?

On average, how many points did those 100 participants take from Participant 2 in this situation?

Participant 1 starts with: 270

Participant 2 starts with: 730

Figure 15: Example: Empirical Expectation



Instructions

Expectations:

For this task, we want to understand your expectations of the other participants.

You will evaluate various situations that were part of the initial task. **Your goal is to guess how the other participants perceived the situation.**

Several cases will be presented. For each case, you have to evaluate participant's entitlement associated to each behavior, from **"very socially inappropriate" (1)** to **"very socially appropriate" (5)**.

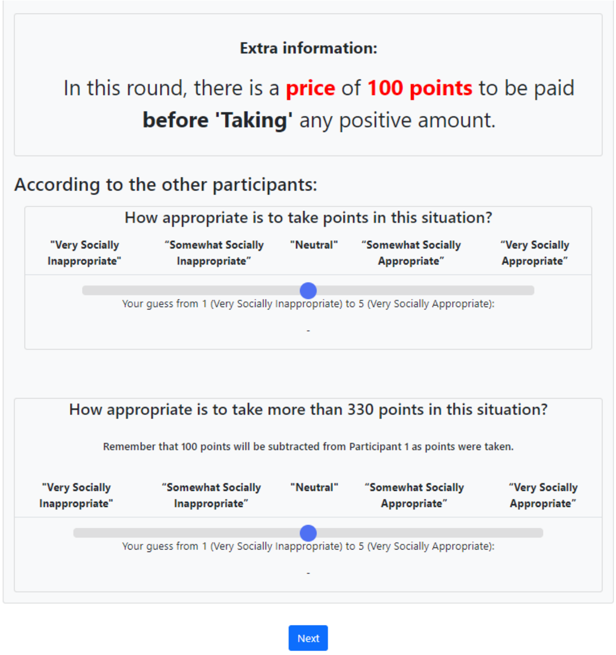
A behavior is appropriate if people most people agree is the "correct" or "ethical" thing to do.

The closer your guess is to the average opinion of the other participants, the greater your gain.

You can earn up to 100 points. 50 points are subtracted from each point your guess is away from the actual number (at most 100 points are subtracted).

One case will be randomly drawn for actual payment.

Figure 16: Information - Normative Expectation



Extra information:

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

According to the other participants:

How appropriate is to take points in this situation?

"Very Socially Inappropriate"
"Somewhat Socially Inappropriate"
"Neutral"
"Somewhat Socially Appropriate"
"Very Socially Appropriate"

Your guess from 1 (Very Socially Inappropriate) to 5 (Very Socially Appropriate):

-

How appropriate is to take more than 330 points in this situation?

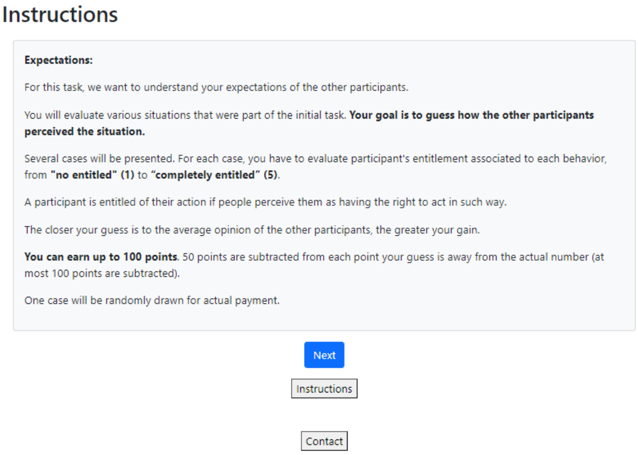
Remember that 100 points will be subtracted from Participant 1 as points were taken.

"Very Socially Inappropriate"
"Somewhat Socially Inappropriate"
"Neutral"
"Somewhat Socially Appropriate"
"Very Socially Appropriate"

Your guess from 1 (Very Socially Inappropriate) to 5 (Very Socially Appropriate):

-

Figure 17: Example: Norm Expectation



Instructions

Expectations:

For this task, we want to understand your expectations of the other participants.

You will evaluate various situations that were part of the initial task. **Your goal is to guess how the other participants perceived the situation.**

Several cases will be presented. For each case, you have to evaluate participant's entitlement associated to each behavior, from **"no entitled" (1)** to **"completely entitled" (5)**.

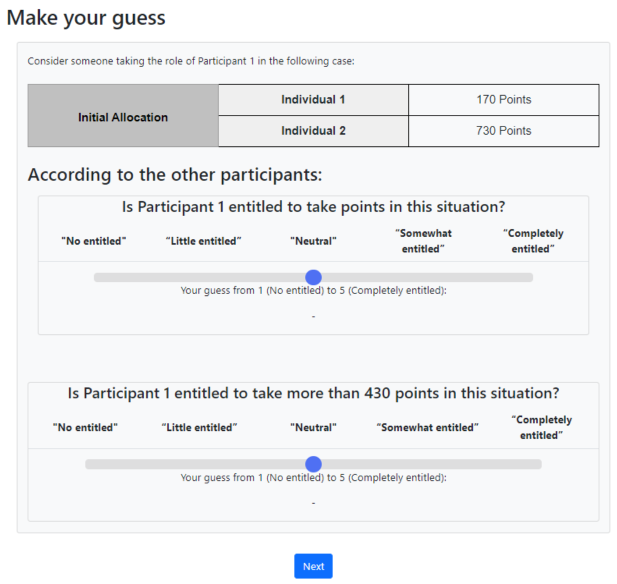
A participant is entitled of their action if people perceive them as having the right to act in such way.

The closer your guess is to the average opinion of the other participants, the greater your gain.

You can earn up to 100 points. 50 points are subtracted from each point your guess is away from the actual number (at most 100 points are subtracted).

One case will be randomly drawn for actual payment.

Figure 18: Information - Entitlement



Make your guess

Consider someone taking the role of Participant 1 in the following case:

Initial Allocation	Individual 1	170 Points
	Individual 2	730 Points

According to the other participants:

Is Participant 1 entitled to take points in this situation?

"No entitled"	"Little entitled"	"Neutral"	"Somewhat entitled"	"Completely entitled"

Your guess from 1 (No entitled) to 5 (Completely entitled):

-

Is Participant 1 entitled to take more than 430 points in this situation?

"No entitled"	"Little entitled"	"Neutral"	"Somewhat entitled"	"Completely entitled"

Your guess from 1 (No entitled) to 5 (Completely entitled):

-

Figure 19: Example: Entitlement