September, 2025 ERI-CES DP 25/04

Deterrence of Unwanted Behavior: a Theoretical and Experimental Investigation

> Penélope Hernández
Universitat de València and ERI-CES, Spain

> Zvika Neeman
Berglas School of Economics, Tel Aviv University, Israel

> Ro'i Zultan
Department of Economics, Ben-Gurion University of the Negev, Israel

eri-ces www.erices.es

Discussion Papers
in Economic Behaviour

Universitat de València

Deterrence of Unwanted Behavior: a Theoretical and Experimental Investigation*

Penélope Hernández† Zvika Neeman‡ Ro'i Zultan§

September, 2025

Abstract

We study whether concentrating enforcement resources - raising sanction probabilities in some areas or times while lowering them elsewhere - reduces socially undesirable behavior. Using a theoretical model and a laboratory experiment, we examine whether splitting a uniform sanction probability p into a higher p_H and a lower p_L can decrease overall violations. Our analysis also provides an empirical test of beneficial splitting, a central concept in the Bayesian persuasion literature, to assess its practical relevance in a parameterized setting.

KEYWORDS: deterrence, compliance, splitting, Bayesian persuasion.

JEL CODES: C72, C91, D80, K42

---

*We thank Ido Erev, Han Bleichrodt, and participants in the Economic Science Association Meeting in Lyon, the Public Economic Theory Meeting in Lyon, Subjective Probability, Utility, and Decision Making in Lucca, and seminar participants in Bonn, Mannheim, and Keio universities for comments and discussion. Neeman acknowledges the support of the Israel Science Foundation, under Grant #1792/23, as well as financial support from the Sapir Center for Economic Development. Zultan acknowledges the support of the Israel Science Foundation, under Grant #524/24. The research carried out in this paper meets the requirements of the Ethics Committee of Tel Aviv University (proposal no. 0009128-2).

†ERI-CES, Universidad de Valencia.

‡Berglas School of Economics, Tel Aviv University.

§Department of Economics, Ben-Gurion University of the Negev.

1 Introduction

A tax compliance officer aims to reduce the under-reporting of income intended to lower tax liabilities. Suppose that allocating enforcement resources (e.g., tax examiners) uniformly across taxpayers yields an audit probability of p for each individual. By concentrating resources, e.g., assigning more examiners to some share λ of taxpayers, however, the audit probability can be raised to $p_H > p$ for those taxpayers at the expense of lowering it to $p_L < p$ for the rest. If taxpayers respond similarly to audit probabilities, a natural question arises: does such a splitting of audit probability help reduce overall under-reporting?

There are two equivalent ways to implement splitting. The direct approach reallocates inspectors across taxpayers, thereby mechanically inducing heterogeneity in audit probabilities. The indirect approach relies on communication: it is based on sending messages that depend on the exogenously determined deployment of tax examiners to shape taxpayers' posterior beliefs about sanction probabilities.[1]

We address this question through a theoretical model and an experiment. The model formalizes the effect of splitting on behavior, while the experiment tests its efficacy in practice. The indirect form of splitting connects to the literature on Bayesian persuasion, which studies how strategic information disclosure influences behavior. Thus, our analysis investigates whether persuasion-based splitting, as well as direct reallocation, can provide practical deterrence benefits in a parameterized setting.

Formally, we study a situation where an individual chooses between a benign action and a socially undesirable one. Examples include reporting income truthfully versus hiding part of it, or parking legally versus illegally. The benign action yields a certain payoff. The undesirable action generates a risky binary

[1] Under the direct approach, $\lambda \in [0, 1]$ taxpayers are audited with probability p_H, and $1 - \lambda$ taxpayers are audited with probability p_L such that $\lambda p_H + (1 - \lambda) p_L = p$ to maintain the same overall expected probability of audit p. Under the indirect approach, a "high chance of audit" message is sent to a fraction $\frac{\lambda p_H}{p}$ of those taxpayers who will be audited, and to a fraction $\frac{\lambda(1-p_H)}{1-p}$ of those who will not, and a "low chance of audit" message is sent to all other taxpayers. Under such a scheme, a fraction $\lambda \in [0, 1]$ of taxpayers receive a "high chance of audit" message, and $1 \geq p_H > p > p_L \geq 0$ can be any two probabilities that satisfy the equation $\lambda p_H + (1 - \lambda) p_L = p$ so as to preserve the same expected probability of audit.

lottery whose outcome depends on whether the individual is sanctioned. Each individual is characterized by a threshold sanction probability above which the benign action is preferred.

We enhance the realism of our model by assuming that the individual's threshold probability is subject to noise and follows a normal distribution. Consequently, the individuals violation function—which relates the likelihood of choosing the socially undesirable action to the probability of sanction—is decreasing and S-shaped with respect to the probability of sanction. Namely, the probability that the individual engages in the unwanted action is initially approximately flat in the probability of sanction, then decreasing, and then approximately flat again as the probability of sanction increases. This S-shaped curve aligns with the intuition that: (i) the likelihood of engaging in the undesirable activity decreases monotonically as the probability of sanction increases; (ii) very low probabilities of sanction have little impact on the individuals behavior; and (iii) high probabilities of sanction are effective deterrents.[2] This intuition is consistent with the result of a well-known experiment that was conducted in Kansas City in 1974 (Kelling et al., 1974), which found that doubling of police patrols had virtually no statistically significant effect on street crime.[3]

Our model is agnostic regarding the individual's underlying decision process. It is consistent with standard decision-making models, including Expected Utility Theory, Prospect Theory, and reliance on small samples. In fact, we intentionally refrain from committing to specific micro-foundations to highlight

[2]All of our results go through if we assume instead that the individual's threshold is deterministic. In this simpler case, the individual's violation function converges to a decreasing step-function with just two steps, as the noise in the individual's threshold decreases to zero. The first step is equal to one (no deterrence) and the second step, which begins at the individual's threshold probability, is equal to zero (complete deterrence).

[3]This finding significantly influenced thinking on deterrence, convincing both academics and police that "police presence does not deter" (Sherman and Weisburd, 1995). Sherman and Weisburd (1995) famously criticized the Kansas City experiment, arguing that Kansas City was too large a unit of analysis for a doubling of patrols to produce an observable effect or for a true reduction in crime to be statistically significant. Sherman and Weisburd repeated the experiment in Minneapolis two decades later, focusing on crime "hot spots" as small as a street corner or city block. They found that doubling police patrols in these hot spots reduced total crime by 6 to 13 percent, while "observed disorder" decreased by half. They concluded that "large increases in dosage may be essential if any effect on crime is to be observed" (Sherman and Weisburd, 1995).

the model's broad applicability.[4]

Having established the violation curve for an individual, we now turn our attention to the population level. A society is made up of many individuals, each characterized by their own violation function. By summing these individual violation functions, we derive an aggregate violation function that relates the probability of sanction to the proportion of the population engaging in the socially undesirable action. This aggregate function retains the same decreasing, S-shaped pattern.[5]

If an individual or an aggregate violation function is decreasing and *convex* throughout its range, then splitting the sanction probability p would increase the overall likelihood of engaging in the unwanted activity. If the violation function is decreasing and *concave* throughout, then splitting the sanction probability p would decrease the overall likelihood of engaging in the unwanted activity. The fact that both single individuals' as well as the aggregate violation functions are decreasing and S-shaped implies that they are first concave, and then convex. This suggests that small values of the sanction probability p may be split in a way that promotes overall deterrence, but large values cannot.

Moreover, for any fixed probability of sanction, decreasing the magnitude of the sanction—increasing the reward or decreasing the penalty from choosing the socially unwanted action—increases the relative attractiveness of the socially unwanted action. Such a change shifts each individual's as well as the aggregate violation function to the right. Accordingly, we say that decreasing the magnitude of the sanction, or increasing the reward/decreasing the penalty from choosing the socially unwanted action, increases the *temptation* to choose the socially unwanted action.

Intuitively, when temptation is very low, the violation function is shifted so

[4]The expected utility model is probably the simplest. In such a model, the individual derives a small benefit from compliance, and has a cost and benefit from noncompliance, depending on whether or not he is sanctioned. Assuming that the benefit from unsanctioned noncompliance is noisy produces a random utility model. Assuming that the individual takes the fraction of times she is sanctioned in $k \geq 1$ times she attempted noncompliance as her estimate of the sanction probability produces a model of reliance on small samples, etc. What is important is that all these models as well as many others would produce a decreasing S-shaped violation function for the individual.

[5]Even if each individual's threshold is deterministic, the aggregate violation remains decreasing and S-shaped so long as the distribution of individuals' thresholds in the population is bell-shaped.

much to the left that it becomes convex on the entire relevant range of sanction probabilities, which in turn implies that splitting increases the violation rate (hurts deterrence). By contrast, when temptation is very high, the violation function is shifted so much to the right that it becomes concave on the entire relevant range, which implies that splitting decreases the rate of violation (improves deterrence). As temptation increases, at some point splitting switches from being detrimental to being beneficial for overall deterrence.[6]

The main theoretical result of this paper formalizes this intuition. We show that for any fixed probability of sanction p, if it is possible to improve individual or aggregate deterrence by splitting p, then it is also possible to improve individual or aggregate deterrence, respectively, by splitting p under higher temptation. And conversely, for any fixed probability of sanction p, if it is impossible to improve individual or aggregate deterrence by splitting p, then it is also impossible to improve individual or aggregate deterrence, respectively, by splitting p under lower temptation.

While the main thrust of our analysis is in line with natural intuitions, it entails two somewhat counterintuitive results. First, the intuitions from violation curves that are always concave or always convex suggest that splitting is beneficial if and only if the violation curve is concave at p. This, however, turns to be false, as splitting is also beneficial at convex regions. Second, splitting becomes beneficial at higher levels of temptations. Accordingly, it may be natural to assume that the benefit from splitting increases monotonically with temptation. This turns to be false as well. Splitting is indeed detrimental below some level of temptation and beneficial above it. However, the absolute benefit from splitting increases as temptation increases above the threshold level up to a point, above which it begins to decline towards zero.

We confirm these theoretical predictions in a laboratory experiment. Our model specifies how individuals respond to the prospect of sanctions, but does not make any assumptions about why individuals respond the way they do. This implies that our model is versatile enough to encompass decisions from experience as well as decisions from description. In our laboratory experiment, we do not provide participants with the sanction probabilities. Instead,

[6]However, it is possible that, as temptation increases, the *absolute* detrimental effect of splitting increases before decreasing and switching to being socially beneficial.

we allow them to learn the probabilities *from experience* by choosing repeatedly between a safe option and a lottery, with feedback following each trial (Barron and Erev, 2003; Hertwig et al., 2004; Hertwig and Erev, 2009; Wulff, Mergenthaler-Canseco, and Hertwig, 2018). This design mirrors real-world applications, where decision-makers lack an explicit description of the sanction probability that is related to violation and must instead infer it from experience. For instance, in our opening example, the chief of police may not even have precise estimates of the probability of sanctions across the different times or areas, and is anyway unlikely to make specific numerical announcements. Rather, potential violators infer the likelihood of sanctions based on their own experiences and, possibly, those of others. Furthermore, the evidence shows that when decision makers have descriptive as well as experiential information, their choices are more similar to choices from experience than to choices from description (Erev et al., 2017; Teoderescu, Amir, and Erev, 2013). Thus, our design allows us to test the main insight that is conveyed by our theoretical analysis in a "realistic" setting. Nonetheless, for completeness, we report in the appendix a supplementary online experiment that confirms our conclusions for decisions based on known probabilities.

The experiment proceeded as follows. In each round of the experiment, participants chose between a safe action, which pays 5 Experimental Currency Units (ECU), and a binary lottery, which pays either a positive or a negative amount. A participant who chooses the safe action is said to be deterred. We implemented splitting by telling subjects to pay attention to a color that is flashed in front of them, because it is related to the probability of receiving the positive payment in the risky binary lottery.[7]

The experimental results both confirm and quantify the observation that splitting of the sanction probability is effective in promoting deterrence in environments in which the temptation to commit the socially unwanted action is

[7]Our main theoretical results pertain to optimal splitting. In practice, calculating the optimal splitting probabilities may be impossible without precise information about the violation curve. Moreover, even if the social planner can determine the optimal splitting probabilities, external constraints may limit the feasible ways to split. We, therefore, show that our theoretical results extend to the problem of choosing whether to split given specific feasible splitting probabilities. The experimental design reflects this situation, fixing the prior-, low- and high-probabilities p, p_L, p_H, respectively, to test the effect of splitting on violation rates with different levels of temptation.

strong and the choice of this action is relatively common—but not in environments in which the temptation to commit the socially unwanted action is weak and the choice of this action is relatively uncommon. These results hold with respect to the overall violation rate at the aggregate level, as well as to the number of individuals from whom splitting promotes deterrence.

Related Literature

The idea that in a game with incomplete information it may be possible to profitably manipulate players' choices through the splitting of prior probabilities dates back to Aumann and Maschler (1995) and is a key observation of the literature on Bayesian persuasion, which originated in Kamenica and Gentzkow (2011). For a recent review of this literature, see Kamenica, Kim, and Zapechelnyuk (2021).

We are aware of only a few experimental studies of Bayesian persuasion. All three papers have a very different focus from ours. Fréchette, Lizzeri, and Perego (2022) study the role of commitment in communication and show that a form of "commitment blindness" leads some senders to overcommunicate when information is verifiable and undercommunicate when it is not (see also Zhou, 2023). Au and Li (2018) perform an experimental study of the relationship between Bayesian persuasion and reciprocity, and Nguyen (2017) studies experimentally whether subjects design their signals in a way that maximizes their expected payoff.

The hotspots literature in criminology (see, e.g., Braga, Papachristos, and Hureau, 2014; and Braga et al., 2019) suggests that enforcement resources should be deployed where they can have the greatest impact. For instance, in a town with two neighborhoods and one police cruiser, this literature explores whether its more effective to assign the cruiser to the first or the second neighborhood. In contrast, we focus on whether it is possible to improve deterrence by concentrating resources in some locations and times, given a fixed amount of resources. Specifically, in a town with two neighborhoods and two police cruisers, is more effective to have both cruisers patrol together, or separately?

Lando and Shavell (2004) and Eeckhout, Persico, and Todd (2010) both considered the question of how to allocate enforcement resources, and argued that it may be beneficial to concentrate enforcement on a subset of the population. Eeckhout, Persico, and Todd also demonstrated this idea empirically using traffic data gathered by the Belgian Police Department. More recently, Hernández and Neeman (2022) have generalized their theoretical results by considering any number of locations, adding uncertainty, considering the question of how to further improve deterrence through Bayesian persuasion, or communication.

The rest of the paper proceeds as follows. In Section 2 we present the model. The experiment is presented in Section 3. Section 4 concludes. Proofs of theoretical results, experimental instructions, as well as the results of the experiment with known probabilities are all relegated to the appendix.

2 Model

2.1 Environments and Violations

We consider the following choice problem. An individual faces a choice between two actions. One is benign, and the other is socially unwanted but beneficial for the individual. Choice of the benign action generates a certain payment to the individual, which we normalize to zero. We refer to the choice of this action as "compliance." Because choice of the socially unwanted action may by subject to sanction, choice of this action induces a risky binary lottery $L(p)$, which is parameterized by the probability of sanction $p \in [0,1]$. With probability p the individual is sanctioned, and the lottery generates a loss $L < 0$ to the individual, and with the complimentary probability $1-p$ the individual is not sanctioned, and the lottery generates a reward $R > 0$. We refer to the choice of this action, which generates an expected payment of $\mathbb{E}[L(p)] = p \cdot L + (1-p) \cdot R$ to the individual, as "committing a violation." Accordingly, individuals who are induced to comply are said to be deterred from committing a violation.

The choice environment we consider is thus characterized by two parameters: a reward $R > 0$, and a loss $L < 0$. Obviously, for any fixed probability of sanction $p \in [0,1]$, increasing the reward R, or decreasing the (absolute value of the) loss $|L|$ strengthens the individual's temptation to commit a violation. In other words, temptation introduces a binary relation over choice environments, which is defined as follows.

Definition 1 *A choice environment* $\langle R, L\rangle$ *induces a stronger temptation to commit a violation than the choice environment* $\langle R', L'\rangle$ *if* $R \geq R'$ *and* $L \geq L'$ *and at least one of these inequalities is strict.*

If a choice environment $\langle R', L'\rangle$ induces a stronger temptation to commit a violation than the choice environment $\langle R, L\rangle$ then we say that "temptation increases" from $\langle R, L\rangle$ to $\langle R', L'\rangle$.

Denote by $\tilde{p}_{R,L}$ the threshold probability of sanction above which the individual prefers to comply in choice environment $\langle R, L\rangle$. That is, when faced with choice problem $\langle R, L, p\rangle$, where R and L denote the Reward and Loss, respectively, and p denotes the probability of sanction, if $p > \tilde{p}_{R,L}$ then the individual prefers to comply; if $p < \tilde{p}_{R,L}$ the individual prefers to commit a violation; and if $p = \tilde{p}_{R,L}$ the individual is indifferent between compliance and commitment of a violation. To simplify notation, as long as it does not cause confusion, we drop the subscript from the threshold sanction and denote it as $\tilde{p}$.

As explained in the introduction, we test our theoretical model experimentally. Because experimental results tend to be subject to noise, the explicit incorporation of noise into the individual's choice through consideration of a random preference model enhances the realism of our model. Specifically, we assume that when faced with a choice environment $\langle R, L\rangle$, the individual's threshold sanction $\tilde{p}$ is distributed according to a normal distribution with mean $\mu_{R,L}$ and standard deviation σ, which we assume to be independent of $\langle R, L\rangle$.[8]

Thus, when faced with a choice problem $\langle R, L, p\rangle$, where $p \in [0, 1]$, the probability that the individual would commit a violation is given by:

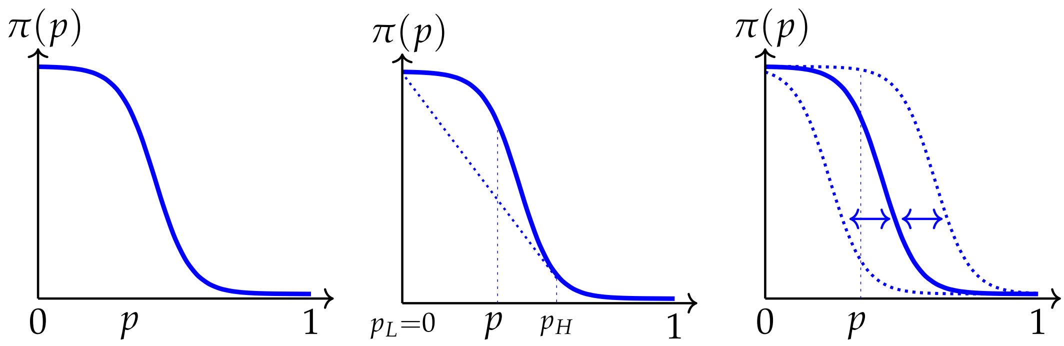
$$\begin{aligned}\pi_{R,L}(p) &\equiv \Pr\left(p \leq \tilde{p}_{R,L}\right) \\ &= 1 - \Phi_{R,L}(p),\end{aligned}$$

where $\Phi_{R,L}$ denotes the cumulative distribution function of a Normal distribution with mean $\mu_{R,L}$ and standard deviation $\sigma_{R,L}$. We refer to the function

[8]Our results continue to hold as long as the standard deviation σ does not decrease too fast in R and L.

$\pi_{R,L}(\cdot)$ as the "violation curve" for environment $\langle R, L\rangle$. As before, to simplify notation, we drop the subscript from the violation curve and denote it as $\pi(\cdot)$.

The function $\Phi_{R,L}$ is an increasing S-shaped function, or a sigmoid function. That is, $\Phi_{R,L}$ is convex and then concave in its argument. Thus, the violation curve $\pi(p)$ is a decreasing S-shaped function of the sanction probability p, which is concave and then convex in p, as depicted in Figure 1a below.



(a) S-shaped violation (b) Optimal splitting (c) Changing fine/reward

Figure 1: The Violation Curve

2.2 Splitting

Splitting the probability of a sanction p into two probabilities $p_L < p < p_H$ may facilitate compliance while maintaining the same amount of enforcement resources. The basic idea is the following. Instead of being faced with the choice problem $\langle R, L, p\rangle$, the individual would be faced with one of two choice problems. With probability λ, the individual would be faced with the choice problem $\langle R, L, p_L\rangle$; and with probability $1-\lambda$, the individual would be faced with the choice problem $\langle R, L, p_H\rangle$. The numbers p_L, p_H and λ are chosen such $0 \leq p_L < p < p_H \leq 1$, and $\lambda p_L + (1-\lambda)p_H = p$. This ensures that the mean probability of a sanction across the two choice problems $\langle R, L, p_L\rangle$ and $\langle R, L, p_H\rangle$, which is given by $\lambda p_L + (1-\lambda)p_H$, remains fixed at p. Being faced with one of two choice problems instead of just with a single choice problem is called splitting because the probability p used in the single choice problem

$\langle R, L, p \rangle$ is split into the probabilities p_L and p_H used in the two choice problems $\langle R, L, p_L \rangle$ and $\langle R, L, p_H \rangle$ in a way that preserves the overall probability that an individual who chooses to commit a violation is sanctioned.

As explained in the introduction, splitting of the prior probability p may be implemented in two different ways. It may be implemented by physically deploying different numbers of enforcement units. Namely, with probability λ deploy number of enforcement units that induce a probability of sanction p_L and with probability $1 - \lambda$ deploy number of enforcement units that induce a probability of sanction p_H, such that $\lambda p_L + (1 - \lambda) p_H = p$. Or splitting may be implemented through communication of information about deployed enforcement resources. Namely, suppose that when enforcement is "on," which occurs with probability p, message "low" is sent with probability $\frac{(1-p_L)(p_H-p)}{(1-p)(p_H-p_L)}$ and message "high" is sent otherwise, and when enforcement is "off" message "low" and is sent with probability $\frac{p_L(p_H-p)}{p(p_H-p_L)}$, and message "high" is sent otherwise. It can be verified that the messages "low" and "high" induce the posterior beliefs p_L and p_H, respectively, with probabilities λ and $1 - \lambda$, respectively.

An individual who is faced with the choice problem $\langle R, L, p_L \rangle$ chooses to commit a violation with probability $\pi(p_L)$; and an individual who is faced with the choice problem $\langle R, L, p_H \rangle$ chooses to commit a violation with probability $\pi(p_H)$. It follows that the expected probability that an individual who is faced with one of the two choice problems $\langle R, L, p_L \rangle$ and $\langle R, L, p_H \rangle$ as described above would commit a violation is equal to:

$$\lambda \pi(p_L) + (1 - \lambda) \pi(p_H).$$

The next definition formalizes the sense in which a probability of a sanction p can be split in a way that promotes overall deterrence.

Definition 2 *A violation curve $\pi(\cdot)$ is said to be* profitably convexifiable *at $p \in (0, 1)$ if there exist two probabilities $p_L < p < p_H$ such that*

$$\lambda \pi(p_L) + (1 - \lambda) \pi(p_H) < \pi(p)$$

for a probability $\lambda \in (0, 1)$ that satisfies $\lambda p_L + (1 - \lambda) p_H = p$.

If a violation curve $\pi(\cdot)$ is profitably convexifiable at p, then there exist two probabilities $p_L < p < p_H$ such that the straight line that connects the points $(p_L, \pi(p_L))$ and $(p_H, \pi(p_H))$ lies below $\pi(\cdot)$ on the interval (p_L, p_H). This is depicted in Figure 1b, where the value of p_L is taken to be equal to zero.

Notably, while a function that is concave on an open interval is profitably convexifiable at any point in this interval, a function can also be profitably convexifiable at points in which it is already convex. Indeed, Figure 1b depicts a violation curve that is both convex and profitably convexifiable at points that are sufficiently close to p_H from below.

If it is possible to split the probability of a sanction in a way that reduces the probability of committing a violation, or that increases compliance, then we say that splitting is socially beneficial. In principle, there could be many pairs of sanction probabilities $p_L < p < p_H$ that make splitting socially beneficial. The pair p_L and p_H that is depicted in Figure 1b is the optimal pair, which maximizes the probability of compliance.

The next Lemma characterizes the range of probabilities that permit socially beneficial splitting.

Lemma 1 *A probability of a sanction p is profitably convexifiable if and only if $p < p^*_{R,L}$ where $p^*_{R,L}$ is given by the unique solution of the problem*

$$\arg\max_{p} \frac{\pi(0) - \pi(p)}{p},$$

provided that $p^*_{R,L} \leq 1$. If $p^*_{R,L} > 1$, then any sanction probability $p < 1$ is profitably convexifiable.

Increasing temptation makes non-compliance relatively more attractive for the individual for every sanction probability. It is therefore natural to assume that increasing temptation increases the rate of violation $\pi(\cdot)$ for every sanction probability $p \in (0, 1)$.

Lemma 2 *Suppose that increasing temptation increases the rate of violation $\pi(\cdot)$ for every sanction probability $p \in (0, 1)$. If temptation increases from choice environment $\langle R, L \rangle$ to $\langle R', L' \rangle$, then $\mu_{R',L'} > \mu_{R,L}$.*

Intuitively, because a larger reward R and a smaller (absolute value of) loss $|L|$ make any lottery $L(p)$ more attractive, the mean $\mu_{R,L}$ is weakly increasing in both R and L.

Recall that we assumed that the standard deviation σ of the individual's threshold $\tilde{p}$ is constant in R and L. Intuitively, we expect that both when temptation is very high and when it is very low, the variance of the individual's choice will be small (because the individual will almost always either violate or comply, respectively). Our model captures this intuition despite σ being constant. This is because when temptation is either very high or very low, Lemma 2 implies that the threshold probability $\tilde{p}$ is almost always either larger or smaller than any $p \in [0, 1]$, respectively. In turn, this implies that the variance of the individual's decision to comply also becomes smaller under very high or very low temptation.

The next proposition describes the main theoretical result of the paper.

Proposition 1 [Increasing temptation shifts the violation curve to the right.] *Suppose that the choice environment $\langle R', L' \rangle$ induces a stronger temptation to commit a violation than choice environment $\langle R, L \rangle$. Then, if the violation curve $\pi(\cdot)$ is profitably convexifiable at sanction probability p in choice environment $\langle R, L \rangle$, then it is also profitably convexifiable at sanction probability p in choice environment $\langle R', L' \rangle$.*

Because the violation curve is decreasing and S-shaped, Proposition 1 implies that increasing temptation shifts the violation curve to the right. As shown in Figure 1c, it follows that the range in which the violation curve is concave expands as temptation increases, and the range in which it is convex expands as temptation decreases. By moving the violation curve sufficiently to the right, it can be made concave over the entire range of sanction probabilities, and by moving the violation curve sufficiently to the left, it can be made convex over the entire range of sanction probabilities.

The following immediate corollary of Proposition 1 provides a testable empirical prediction of our main result.

Corollary 1 *Suppose the choice environment $\langle R', L' \rangle$ induces a stronger temptation to commit a violation than the choice environment $\langle R, L \rangle$. If a sanction probability p can be profitably convexified by splitting it into p_L and p_H in environment $\langle R, L \rangle$,* *then p can also be profitably convexified by splitting it into p_L and p_H in environment $\langle R', L' \rangle$.*[9]

Proposition 1 is formulated for the case of a single individual. By aggregating individuals' violation functions, it is possible to obtain an aggregate analog of Proposition 1 as follows.

Suppose that there are n different individuals. Let $\tilde{p}^i_{R,L}$ denote the threshold probability of sanction above which individual i prefers to comply in choice environment $\langle R, L \rangle$. Suppose that individuals compliance decisions are stochastically independent. That is, when faced with choice problem $\langle R, L, p \rangle$, if $p > \tilde{p}^i_{R,L}$ then individual i complies; if $p < \tilde{p}^i_{R,L}$ then individual i commits a violation; and if $p = \tilde{p}^i_{R,L}$ then individual i is indifferent between compliance and the commitment of a violation, independently of whether other individuals' comply or not. As before, to simplify notation, we drop the subscript from the threshold sanction and denote it as $\tilde{p}^i$.

Each individual i's threshold sanction $\tilde{p}$ is normally distributed, with mean $\mu^i_{R,L}$ and standard deviation $\sigma^i_{R,L}$. We denote individual i's distribution by $\Phi^i_{R,L}(p)$. Thus, when faced with a choice problem $\langle R, L, p \rangle$, the mean fraction of individuals who would commit a violation is given by:

$$\begin{aligned} \pi^{\Sigma}_{R,L}(p) &\equiv \frac{1}{n} \sum_{i=1}^{n} \Pr\left(p \leq \tilde{p}^i_{R,L}\right) \\ &= 1 - \Phi^{\Sigma}_{R,L}(p), \end{aligned}$$

where $\Phi^{\Sigma}_{R,L}$ denotes the cumulative distribution function of a Normal distribution with mean $\mu^{\Sigma}_{R,L} = \frac{1}{n} \sum_{i=1}^{n} \mu^i_{R,L}$ and standard deviation $\sigma^{\Sigma}_{R,L} = \sqrt{\frac{1}{n^2} \sum_{i=1}^{n} (\sigma^i_{R,L})^2}$. We refer to the function $\pi^{\Sigma}_{R,L}(\cdot)$ as the "aggregate violation curve" for environment $\langle R, L \rangle$. As before, to simplify notation, we drop the subscript from the violation curve and denote it as $\pi^{\Sigma}(\cdot)$. Because it is equal to a sum of decreasing S-shaped functions, the aggregate violation curve $\pi^{\Sigma}(p)$ is a also a decreasing S-shaped function of the sanction probability p, which is concave and then convex in p, as depicted in Figure 1a above.

[9] However, the benefit from splitting p may be lower in environment $\langle R', L' \rangle$ compared to environment $\langle R, L \rangle$.

We thus have the following proposition, which describes the aggregate analog to Proposition 1.

Proposition 2 *Suppose that the choice environment $\langle R, L\rangle$ induces a stronger temptation to commit a violation than choice environment $\langle R', L'\rangle$. Then, if the aggregate violation curve $\pi^{\Sigma}(\cdot)$ (which measures the mean fraction of individuals who violate) is profitably convexifiable at sanction probability p in choice environment $\langle R', L'\rangle$, then it is also profitably convexifiable at sanction probability p in choice environment $\langle R, L\rangle$.*

The proof of Proposition 2 is similar to that of Proposition 1 and is omitted. The following corollary provides a testable empirical prediction of aggregate behavior.

Corollary 2 *Suppose that the choice environment $\langle R, L\rangle$ induces a stronger temptation to commit a violation than choice environment $\langle R', L'\rangle$. Then, the number of individuals whose violation curve $\pi^{i}(\cdot)$ is profitably convexifiable at sanction probability p through splitting into probabilities p_L and p_H in choice environment $\langle R, L\rangle$ is larger than or equal to the number of individuals whose violation curve is profitably convexifiable at p through splitting into probabilities p_L and p_H in choice environment $\langle R', L'\rangle$.*

3 Experiment

To test the benefit of splitting under different levels of temptation, we ran an experiment in which participants repeatedly chose between compliance and violation. Each participant made 100 choices under a pooling condition and 100 choices under a splitting condition. We manipulated temptation by changing the reward and loss values between subjects. The probabilities of sanction p, p_L, p_H were fixed across all participants and trials.

3.1 Design and Procedure

The experiment consisted of one block of 100 splitting trials and one block of 100 pooling trials, with the order of blocks randomized at the individual level. In each trial, each participant observed a signal in the shape of a colored circle and chose between a safe option (comply) and a risky option (violate). The safe option always yielded a payoff of 5 ECU (Experimental Currency Units). The risky option yielded a lottery between a reward—capturing the benefit from committing a violation without being sanctioned—and a loss—capturing the sanction. The reward was always higher, and the loss was always lower than 5 ECU. These values remained fixed across all 200 trials for any individual but varied between individuals depending on the treatment. After making their choice, participants observed their realized payoff in the trial and proceeded to the next trial.

The probability of obtaining the reward (and otherwise the loss) could vary from trial to trial depending on the color of the circle. In the pooling trials, the circle was always yellow, and the probability of sanction upon choosing to commit a violation was fixed at $p = \frac{3}{5}$. In the splitting trials, the circle was blue with probability $\lambda = .44$ and red with probability $1 - \lambda = .56$. In the "red" trials, the probability of being sanctioned was $p_H = 1$, i.e., this option yielded a loss with certainty. In the "blue" trials, the risky option yielded the loss with a probability of $p_L = \frac{1}{12}$. The mean probability of a sanction in the splitting trials was thus $.56 \cdot 1 + .44 \cdot \frac{1}{12} = .597$, slightly less than in the pooling trials. The instructions stated that a red circle indicated a sure loss while leaving all other probabilities unspecified, reflecting the realistic assumption that while certainties are identifiable, exact probabilities are seldom known.

We ran five between-subjects treatments in which we varied the reward and loss parameters, as summarized in Table 1. The treatments were Weak ($R = 10, L = -50$), Medium-weak ($R = 10, L = -30$), Medium ($R = 30, L = -30$), Medium-strong ($R = 30, L = -10$), and Strong temptation ($R = 50, L = -10$).[10] The treatments manipulated temptation by alternately varying the loss and reward payments, with the reward increasing from Medium-weak to Medium and from Medium-strong to Strong temptation whereas the (absolute) loss decreased from Weak to Medium-weak and from Medium to Medium-strong temptation.

The experiment was carried out at the Laboratory for Research in Experimental Economics (LINEEX) in the University of Valencia in December, 2022,

[10] In choosing the parameters for our experiment, we have relied on the estimates produced by Erev et al. (2017)'s *Best Estimate and Sampling Tools* (BEAST) model of choice under uncertainty. Accordingly, our findings both rely on and qualitatively validate the theoretical predictions produced by the BEAST model.

Table 1: Experimental Treatments

Temptation	Safe	Reward	Loss
Weak	5	10	-50
Medium-weak	5	10	-30
Medium	5	30	-30
Medium-strong	5	30	-10
Strong	5	50	-10

and March, 2023. Participants were recruited from the subject pool of the laboratory using HRoot (Bock, Baetge, and Nicklisch, 2014). For each treatment, we ran one session with 50 participants for a total of 250 across the five treatments. The instructions, available in the appendix, were read aloud at the beginning of each session/treatment. Each session lasted approximately 50 minutes. The average payment to participants was 13.8 Euros.

3.2 Hypotheses

Because participants in the experiment faced three different probabilities of sanction: $\frac{1}{12}$, $\frac{3}{5}$, and 1, each treatment of the experiment generated three points on the participants' violation curve. We refer to the piecewise linear curve resulting from connecting these three points as the *experimental violation curve*. Whether splitting is beneficial is directly tied to the curvature of this curve. If the line that connects the points with sanction probabilities $p = \frac{1}{12}$ and $p = 1$ lies below the point with sanction probability $p = \frac{3}{5}$, then the sanction probability $p = \frac{3}{5}$ is profitably convexifiable through splitting to $p_L = \frac{1}{12}$ and $p_H = 1$.

Our first hypothesis tests Corollary 1.

Hypothesis 1 *If the experimental violation curve is concave for some treatment, then it is also concave for all treatments that induce a higher temptation.*

Our second hypothesis tests Corollary 2.

Hypothesis 2 *As temptation increases, the violation curves are concave for more individuals.*

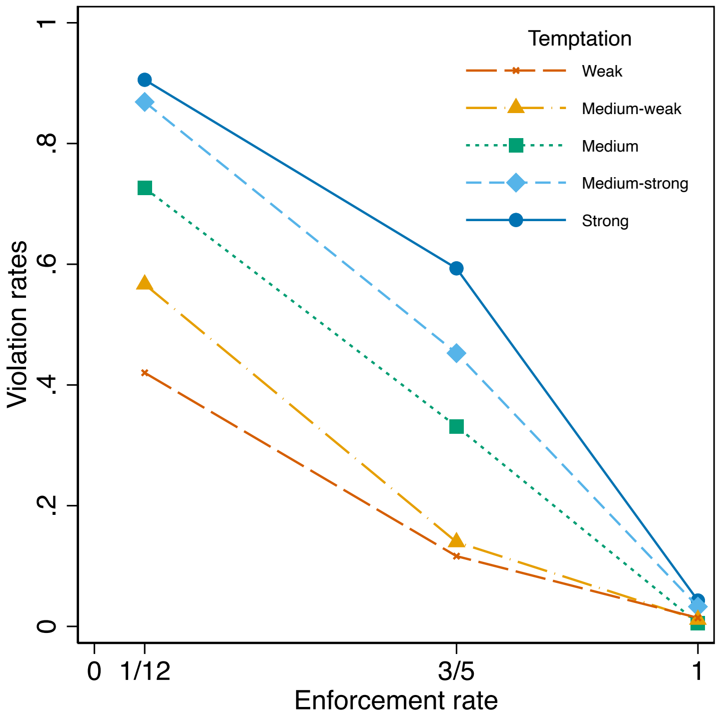


Figure 2: Violation rates across treatments.

3.3 Results

We first analyze the effects of temptation on the concavity of the experimental violation curve and the effectiveness of splitting at the aggregate level. We proceed with analyses and tests at the individual level.[11] The replication materials are available at https://osf.io/a86qm/?view_only=5045a91d030340daad217351aeeef397.

3.3.1 Aggregate Violation

Figure 2 depicts the aggregate violation rates by treatment and probability of sanction. The experimental violation curve is approximately linear for medium temptation, concave for higher temptations and convex for lower temptation, in line with Hypothesis 1.

To formally measure the curvature of the experimental violation curves, we calculated the violation rate of each participant for each enforcement rate.[12] We then estimated the following regression of individual violation rates on treatment dummies (M = Medium Temptation; MW = Medium-weak Temptation; MS = Medium-strong Temptation; S = Strong Temptation) and their interactions with the enforcement rate and the squared enforcement rate, clustering robust standard errors on participants:

$$\begin{aligned} VR_{ip} = \beta_0 + \beta_1 p + \beta_2 p \times p \\ + \beta_3 MW_i + \beta_4 MW_i \times p + \beta_5 MW_i \times p \times p \\ + \beta_6 M_i + \beta_7 M_i \times p + \beta_8 M_i \times p \times p \\ + \beta_9 MS_i + \beta_1 0 MS_i \times p + \beta_1 1 MS_i \times p \times p \\ + \beta_1 2 S_i + \beta_1 3 S_i \times p + \beta_1 4 S_i \times p \times p + \epsilon_{ip}. \end{aligned}$$

Table 2 presents terms based on the regression coefficients that capture the convexity (if positive) and concavity (if negative) of the experimental violation curve across the five treatments (see the appendix for the full regression table). The results support the observation based on Figure 2. Namely, the experimental violation curve is significantly convex under Weak and Medium-weak temptation. It is indistinguishable from linear under Medium temptation, and significantly concave under Medium-strong and Strong temptation (at the 10% level for the former).[13]

[11] All of the results remain virtually unchanged when restricting the analysis to the last 50 of the 100 trials. Similarly, the regression presented in Table 3 yields similar results when including different time trends by treatments. The overall time trends from this specification indicates that, on average, the probability of choosing to comply increases by 7 precentage points across the 100 trials.

[12] Each of the 250 participants was faced with three sanction probabilities, resulting in five treatments and 750 observations overall.

[13] Note that concavity does not increase monotonically with temptation as the experimental

Table 2: Curvature of the violation curves.

	Term	Convexity	t-value	p-value
Weak	β_2	0.357	2.98	0.003
Medium-weak	$\beta_2 + \beta_5$	0.544	6.39	0.000
Medium	$\beta_2 + \beta_8$	-0.051	-0.36	0.722
Medium-strong	$\beta_2 + \beta_{11}$	-0.269	-1.65	0.099
Strong	$\beta_2 + \beta_{14}$	-0.843	-4.43	0.000

Notes: Marginal self interactions of enforcement rate based on an OLS regression of subject-level mean violation rate by enforcement rate with robust standard errors clustered on participants.

Table 3: Regression on violation rates.

	Coefficient	Robust std. error	z-statistic	p-value
Medium-weak	0.210	0.213	0.99	.324
Medium	1.325	0.216	6.15	.000
Medium-strong	1.837	0.220	8.34	.000
Strong	2.404	0.239	10.04	.000
Split	0.624	0.1997	3.16	.002
Medium-weak × Split	0.139	0.227	0.61	.541
Medium × Split	-0.652	0.240	-2.72	.007
Medium-strong × Split	-0.869	0.239	-3.64	.000
Strong × Split	-1.297	0.254	-5.11	.000
Constant	-2.027	0.168	-12.10	.000

Notes: Logistic regression of violation rates on treatment interacted with splitting with robust standard errors clustered on participants.

Table 3 presents the results of a logistic regression of violation on treatment interacted with splitting with robust standard errors clustered on participants. Figure 3 plots the results from the regression, with the predicted violation rate presented in the left panel and the estimated marginal effect of splitting presented in the right panel. The regression supports Hypothesis 1. Splitting of the sanction probability $\frac{3}{5}$ into the sanction probabilities $\frac{1}{12}$ and 1 reduces violations significantly under Strong temptation ($z = 4.32, p < .001$) and weakly significantly under Medium-strong temptation ($z = 1.80, p = .072$); has no significant effect under medium temptation ($z = 0.21, p = .837$); and significantly increases violations under Medium-weak ($z = 6.71, p < .001$) and Weak temptation ($z = 3.14, p = .002$).[14]

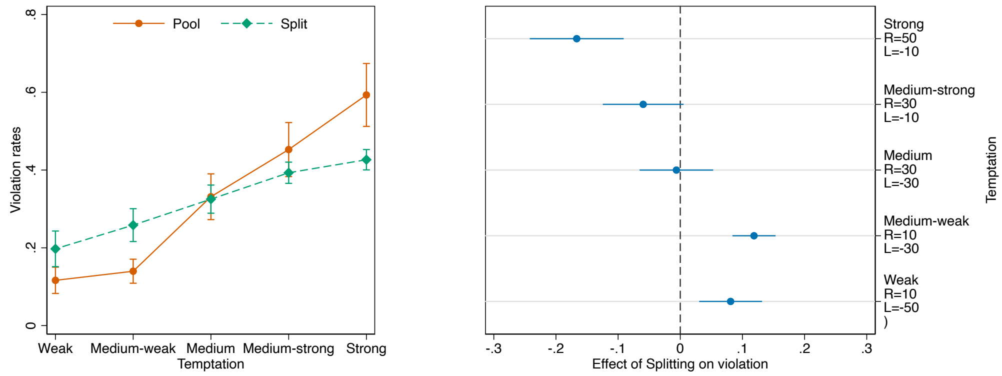
A pairwise comparison of the effect of splitting on the violation rate between adjacent treatments reveals that the treatment effects arise from changes in the reward R rather than in the loss L. Splitting is less detrimental (more beneficial) when increasing R from Medium-weak to Medium and from Medium-strong to Strong temptation ($z = 3.57, p < .001$ and $z = 2.10, p = .035$, respectively). In contrast, increasing L from Weak to Medium-weak or from Medium to Medium-strong temptation has no significant effect ($z = -1.20, p = .229$ and $z = 1.19, p = .233$, respectively).

Result 1 *the results strongly support Hypothesis 1. Splitting is socially beneficial for high temptation and socially detrimental for low temptation. The benefit from splitting with our parameterization is sensitive to the reward R but not to the loss L.*[15]

violation curve is more concave under Weak compared to Medium-weak temptation. Although at first sight this appears to contradict the intuition behind our prediction, it is consistent with our theoretical results, which only show that *convexifiability* and not concavity is monotonically increasing with temptation.

[14]Wilcoxon signed-rank tests comparing individual-level violation rates between splitting and pooling in each treatment yield identical results.

[15]We are nonetheless hesitant to draw any general conclusion based on only four comparisons and believe that this pattern may not be generalizable. To see this, consider a treatment with $R = 10$ and $L = -10$. It is plausible that behavior in this treatment would be similar to behavior in the Medium temptation treatment, where $R = 30$ and $L = -30$. That is, a change in the reward (compared to the Medium-strong treatment) would have no effect, while a change in the loss (compared to the Medium-weak treatment) would.



(a) Violation rates.

(b) Marginal effects.

Figure 3: Effects of splitting enforcement on violation rates.

3.3.2 Individual Heterogeneity

To test Hypothesis 2, we calculated for each participant the absolute difference in violation rate between the splitting and pooling trials. Figure 4 presents the share of individuals for whom splitting is beneficial or detrimental (i.e., reduces or increases violations, respectively) by treatment. Significance is based on t-tests at the individual level. The share of individuals for whom splitting is beneficial increases with temptation, while the share of individuals for whom splitting is detrimental mostly decreases with temptation. The exception is that splitting is detrimental for more individuals under Medium-weak compared to Weak temptation. This deviation from monotonicity is, perhaps, not surprising considering that our model predicts that the violation curve flattens if shifted enough to the left or to the right. The fact that compliance is the modal choice even with $p = \frac{1}{12}$ under Weak temptation suggests that this is the case. Furthermore, this non-monotonicity disappears when excluding nine participants who never violated in the Weak and Medium-weak treatments.

An ordered logistic regression of the five types presented in Figure 4 on treatment reveals significant effects in the predicted direction. Mirroring our results for beneficial splitting, the shift in the type distribution is significant when increasing R from Medium-weak to Medium temptation and from Medium-strong to Strong temptation ($z = 3.07, p = .002$ and $z = 2.70, p = .007$, respec-

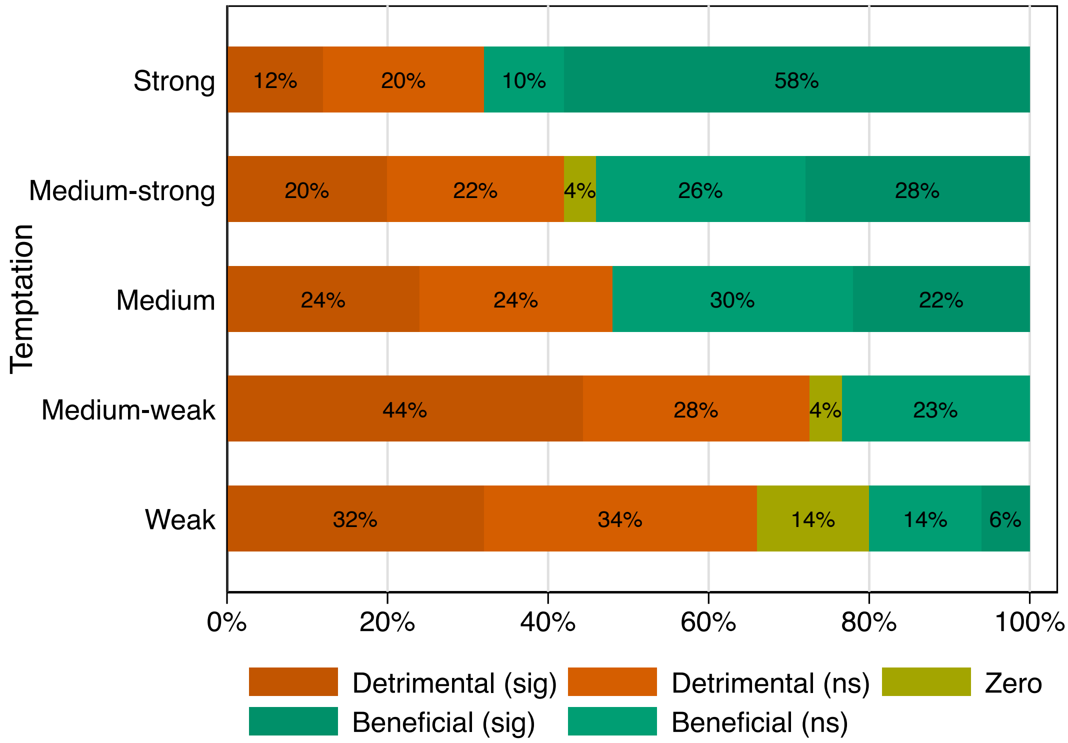


Figure 4: Individual heterogeneity.

tively) but not when increasing L from Weak to Medium-weak or from Medium to Medium-strong temptation ($z = -0.98, p = .329$ and $z = 0.65, p = .513$, respectively).

Result 2 *The individual-level analysis supports Hypothesis 2. The share of individuals for whom splitting is socially beneficial weakly increases with temptation. This share is sensitive to the reward R but not to the loss L.*

4 Conclusion

The importance of the experiment lies in that it allows us to quantify exactly how high temptation needs to be in order for splitting to be effective in promoting overall deterrence. In our experiment, the sanction probability $p = 0.6$

can be split in a way that promotes overall deterrence if the binary lottery pays -10 and 50 ECU, when the individual is and is not sanctioned, respectively. In this environment, not splitting the sanction probability $p = 0.6$ implies that 59% of participants' choices are for the socially unwanted action. If the binary lottery pays -30 and $+30$ ECU upon sanction and no sanction, respectively, then splitting has no effect on overall deterrence. In such an environment, not splitting the sanction probability $p = 0.6$ implies that 33% of participants' choices are for the socially unwanted action. Finally, if the binary lottery pays -50 and $+10$ ECU upon sanction and no sanction, respectively, then splitting hurts overall deterrence. In such an environment, not splitting the sanction probability $p = 0.6$ implies that 12% of participants' choices are for the socially unwanted action.

To help put the numbers above in perspective, note that Slemrod (2019) reports that while the overall non compliance rate of IRS tax liabilities is 16.3 percent, when there is little to no third-party-reported information (as in the case of self-employment income), the estimated noncompliance rate [of IRS payments] increases to 63 percent. This suggests that splitting the probability of audit on self-employed tax payers for whom there is little third-party reported information would improve overall rates of compliance.

References

Au, Pak Hung and King King Li (2018). Bayesian persuasion and reciprocity: theory and experiment. *Available at SSRN 3191203*.

Aumann, Robert J and Michael Maschler (1995). Repeated games with incomplete information. MIT Press.

Barron, Greg and Ido Erev (2003). Small feedback-based decisions and their limited correspondence to description-based decisions. *Journal of behavioral decision making* 16(3), pp. 215–233.

Bock, Olaf, Ingmar Baetge, and Andreas Nicklisch (2014). Hroot: hamburg registration and organization online tool. *European Economic Review* 71, pp. 117–120.

Braga, Anthony A, Andrew V Papachristos, and David M Hureau (2014). The effects of hot spots policing on crime: an updated systematic review and meta-analysis. *Justice quarterly* 31(4), pp. 633–663.

Braga, Anthony A, Brandon S Turchan, Andrew V Papachristos, and David M Hureau (2019). Hot spots policing and crime reduction: an update of an ongoing systematic review and meta-analysis. *Journal of experimental criminology* 15, pp. 289–311.

Eeckhout, Jan, Nicola Persico, and Petra E Todd (2010). A theory of optimal random crackdowns. *American Economic Review* 100, pp. 1104–1135.

Erev, Ido, Eyal Ert, Orly Plonsky, Dana Cohen, and Or Cohen (2017). From anomalies to forecasts: toward a descriptive model of decisions under risk, under ambiguity, and from experience. *Psychological Review* 124(4), pp. 369–409.

Fréchette, Guillaume R, Alessandro Lizzeri, and Jacopo Perego (2022). Rules and commitment in communication: an experimental analysis. *Econometrica* 90(5), pp. 2283–2318.

Hernández, Penélope and Zvika Neeman (2022). How bayesian persuasion can help reduce illegal parking and other socially undesirable behavior. *American Economic Journal: Microeconomics* 14(1), pp. 186–215.

Hertwig, Ralph, Greg Barron, Elke U Weber, and Ido Erev (2004). Decisions from experience and the effect of rare events in risky choice. *Psychological science* 15(8), pp. 534–539.

Hertwig, Ralph and Ido Erev (2009). The description–experience gap in risky choice. *Trends in cognitive sciences* 13(12), pp. 517–523.

Kamenica, Emir and Matthew Gentzkow (2011). Bayesian persuasion. *American Economic Review* 101, pp. 2590–2615.

Kamenica, Emir, Kyungmin Kim, and Andriy Zapechelnyuk (2021). Bayesian persuasion and information design: perspectives and open issues. *Economic Theory* 72(3), pp. 701–704.

Kelling, George, Anne M Pate, Dennis Dieckman, and Christine Brown (1974). The kansas city preventive patrol experiment: technical report. Tech. rep. Washington DC: Police Foundation.

Lando, Henrik and Steven Shavell (2004). The advantage of focusing law enforcement effort. *International Review of Law and Economics* 24, pp. 209–218.

Nguyen, Quyen (2017). Bayesian persuasion: evidence from the laboratory. *Work. Pap., Utah State Univ., Logan*.

Sherman, Lawrence W and David Weisburd (1995). General deterrent effect of police patrol in crime "hot spots": a randomized controlled trial. *Justice Quarterly* 12(4), pp. 625–648.

Slemrod, Joel (2019). Tax compliance and enforcement. *Journal of Economic Literature* 57(4), pp. 904–954.

Teoderescu, Kinneret, Michal Amir, and Ido Erev (2013). The experience–description gap and the role of the inter decision interval. In: *Decision making: neural and behavioural approaches*. Ed. by V. S. Chandrasekhar Pammi and Narayanan Srinivasan. Vol. 202. Progress in Brain Research. Elsevier. Chap. 6, pp. 99–115.

Wulff, Dirk U, Max Mergenthaler-Canseco, and Ralph Hertwig (2018). A meta-analytic review of two modes of learning and the description-experience gap. *Psychological bulletin* 144(2), p. 140.

Zhou, Junya (2023). Costly verification and commitment in persuasion. *Journal of Economic Behavior & Organization* 212, pp. 1100–1142. ISSN: 0167-2681.

Appendix A Proofs

A.1 Proof of Lemma 1

The first order condition of the optimization problem yields

$$\pi(p^*) = \pi(0) + \pi'(p^*)p^*. \tag{1}$$

The solution of Equation (1) describes the unique probability of a sanction p^* that has the property that the tangent of the violation curve $\pi(p)$ at p^* is such that: (i) the tangent lies below $\pi(p)$ on the interval $(0, p^*)$; and (ii) the tangent intersects the violation curve at the point $(0, \pi(0))$. It follows that all the probabilities $p < p^*$ are profitably convexifiable, and no probability $p > p^*$ is profitably convexifiable.

A.2 Proof of Lemma 2

Notice that if temptation increases from choice environment $\langle R, L\rangle$ to $\langle R', L'\rangle$, then $\pi_{R',L'}(p) \geq \pi_{R,L}(p)$ for all $p \in [0, 1]$.

We express the function $\pi(.)$ by using the error function $erf(\cdot)$ as follows:

$$\pi_{R,L}(p) = \frac{1}{2} - \frac{1}{\sqrt{\pi}} \int_0^{\frac{p-\mu}{\sigma\sqrt{2}}} e^{\frac{-t^2}{2}} dt$$

Writing the above inequality for $p = \mu$, we get:

$$\frac{1}{2} - \frac{1}{\sqrt{\pi}} \int_0^{\frac{\mu-\mu'}{\sigma'\sqrt{2}}} e^{\frac{-t^2}{2}} dt \geq \frac{1}{2} - \frac{1}{\sqrt{\pi}} \int_0^{\frac{\mu-\mu}{\sigma\sqrt{2}}} e^{\frac{-t^2}{2}} dt$$

or

$$-\frac{1}{\sqrt{\pi}} \int_0^{\frac{\mu-\mu'}{\sigma'\sqrt{2}}} e^{\frac{-t^2}{2}} dt \geq 0$$

Consequently, $\mu' \geq \mu$ and the condition holds.

A.3 Proof of Proposition 1

As the choice environment $\langle R, L\rangle$ induces a stronger temptation to commit a violation than the choice environment $\langle R', L'\rangle$ then $\mu \geq \mu'$ and $\sigma = \sigma'$. Notice that the violation curve $\pi_{R,L}$ under environment $\langle R, L\rangle$ is the violation curve $\pi_{R',L'}$ under environment $\langle R', L'\rangle$ shifts to the left in a parallel way. Consequently,

$$\pi_{R',L'}(p) = \pi_{R,L}(p + (\mu - \mu')) \tag{2}$$

for all $p \in \mathbf{R}$.

Let $p^*_{R',L'}$ be the solution of 1 and $\tilde{p} < 0$ such that $\pi_{R,L}(\tilde{p}) + (\mu - \mu') = \pi_{R',L'}(0)$. As the line that connect $(0, \pi_{R',L'}(0))$ and $(p^*_{R',L'}, \pi_{R',L'}(p^*_{R',L'}))$ is tangent at $p^*_{R',L'}$, then the line that connect $(\tilde{p}, \pi_{R',L'}(\tilde{p}))$ and $(p^*_{R',L'}, \pi_{R',L'}(p^*_{R',L'}))$ is also bellow than $\pi_{R',L'}$.

Let $p \in [0, p^*_{R',L'}]$ profitably convexifiable, then we can write $p = \lambda_{R',L'} p^*_{R',L'}$ for $\lambda_{R',L'} \in [0, 1]$ and

$$(1 - \lambda_{R',L'})\pi_{R',L'}(0) + \lambda_{R',L'}\pi_{R',L'}(p^*_{R',L'}) \leq \pi_{R',L'}(\lambda_{R',L'} p^*_{R',L'}) = \pi_{R',L'}(p)$$

Let us check that p is also convexifiable under the environment $\langle R, L\rangle$ by splitting at 0 and $p^*_{R',L'} + (\mu - \mu')$. Let $\tilde{\lambda}$ be such that $p = (1 - \tilde{\lambda})0 + \tilde{\lambda}(p^*_{R',L'} + (\mu - \mu'))$.

$$\begin{aligned}
(1-\tilde{\lambda})\pi_{R,L}(0) + \tilde{\lambda}\pi_{R,L}(p^*_{R',L'} + (\mu - \mu')) &= (1-\tilde{\lambda})\pi_{R,L}(0) + \tilde{\lambda}\pi_{R',L'}(p^*_{R',L'}) \\
&< \pi_{R',L'}((1-\tilde{\lambda})(-(\mu - \mu')) + \tilde{\lambda}(p^*_{R',L'})) \\
&= \pi_{R,L}(\tilde{\lambda}(p^*_{R',L'}) + (\mu - \mu')) \\
&= \pi_{R,L}(p)
\end{aligned}$$

where the first and third equality come from the equation 2 and the second inequality holds because the line that connect $(-(\mu - \mu')), \pi'(-(\mu - \mu'))$ with $(p^*_{R',L'}, \pi'(p^*_{R',L'}))$ is below to the tangent at $p^*_{R',L'}$.

Appendix B Regression table

Table B.1: Regressions for Table 2.

	Violation rate
Constant	0.487***
	(7.93)
p	-0.836***
	(-4.80)
p×p	0.363**
	(2.98)
Medium-weak	0.176*
	(2.11)
Medium-weak×p	-0.367
	(-1.68)
Medium-weak×p×p	0.188
	(1.27)
Medium	0.300***
	(3.81)
Medium×p	0.109
	(0.45)
Medium×p×p	-0.418*
	(-2.23)
Medium-strong	0.435***
	(6.24)
Medium-strong×p	0.213
	(0.86)
Medium-strong×p×p	-0.630**
	(-3.12)
Strong	0.427***
	(6.55)
Strong×p	0.806**
	(3.07)
Strong×p×p	-1.204***
	(-5.38)
Observations	49,869

Notes: OLS regression of individual violation rates by enforcement rate p with robust standard errors clustered on individuals. t-statistics in parentheses. * ($p<0.05$), ** ($p<0.01$), *** ($p<0.001$)

Appendix C Experimental instructions

Medium temptation

SCREEN 1: WELCOME

Welcome and thank you for participating in this experiment. Please read the instructions very carefully; the better you understand the instructions, the more money you can earn. During the experiment, all monetary amounts will be in ECU (Experimental Currency Unit). Your earnings during the experiment will be converted to euros at the end of the experiment.

The exchange rate is 100 ECU = 1 euro.

In this experiment you will not interact with other participants. Your earnings will depend on your decisions and chance, as explained below.

After the experiment, we will ask you to complete a brief questionnaire. The questionnaire data, as well as all your decisions during the experiment, will be anonymous.

SCREEN 2: THE TASK

This experiment has 2 blocks.

In each block, you can click on two buttons a total of 100 times. Each time you can choose which button to click. After clicking, you will see your payment for choosing that button. Your earnings will be the sum of your payments from the 100 clicks in each of the 2 blocks.

One button will give you a reward of 5 ECUs every time you click it. The other button will give you 30 or -30 ECUs.

Above the buttons, you will see a circle. The color of the circle may change from click to click. Pay attention, as the color of the circle will determine the probability of obtaining one payment or another on the 30/-30 button.

In one block, you will always see the yellow color. This means that if you choose the 30/-30 button, sometimes you will get -30 ECUs and other times 30 ECUs.

In the other block, the circle can be red or blue.

- The red color indicates that if you choose the 30/-30 button, you will always get -30.
- The blue color indicates that if you choose the 30/-30 button, sometimes you will get -30 ECUs and other times 30 ECUs.

Screen 3: Beginning of Block

Remember:

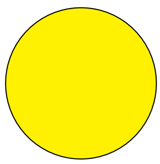
In one block, you will always see the yellow color. This means that if you choose the 30/-30 button, sometimes you will get -30 ECUs and other times 30 ECUs.

In the other block, the circle can be red or blue.

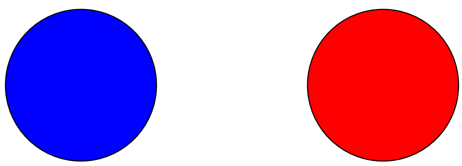
- The red color indicates that if you choose the 30/-30 button, you will always get -30.
- The blue color indicates that if you choose the 30/-30 button, sometimes you will get -30 ECUs and other times 30 ECUs.

The following colors will appear during this block:

<Pooling block>

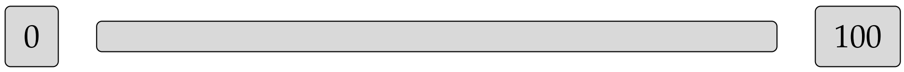


<Splitting block>



Screen 4: Risk elicitation task

There is a hidden mine in one of the numbers between 1 and 100. You can choose a number between 1 and 100. If the mine is in a number less or equal to the number that you have chosen, then it explodes to you and you do not win anything. If the mine is in a number greater than the number that you have chosen, then it does not burst to you and your payment is your chosen number.



Click and slide on the top bar to select the desired number.

Appendix D Online experiment

This appendix describes the online experiment studying decisions from description. The pre-registration is available at https://osf.io/6x2ze/?view_only=08ff7722d3bd4b9ab15c7215e7403e04. The experimental screens and replication files are available at https://osf.io/jh63z/?view_only=5298ab3388a74da9aeb63ee919cc4f75.

D.1 Experimental design and procedure

Participants. The experiment was conducted on Prolific with participants recruited from the United Kingdom. As pre-registered, we recruited participants incrementally until reaching 100 valid participants, as described below. The average duration taken to complete the experiment was 5:06 minutes, and the median duration was 4:12 minutes. Payments ranged between 1.00 GBP and 2.00 GBP with a mean of 1.62 GBP and a median of 1.55 GBP.

Procedure. Participants received 1.50 GBP and chose between a safe option of and additional 0.05 GBP and a lottery in which they can receive a reward or incur a loss. Each participant made ten choices, with one random choice determining their payoff. The values of the reward and loss were as in the laboratory experiments, stated in pennies. For example, in the Weak-temptation treatment, the lottery paid 0.10 GBP or -0.50 GBP. The probabilities of incuring the loss were either $p = .6$ or $p_L = \frac{1}{12}$. We assume that everyone complies under a certain sanctioning probability of $p_H = 1$.[16]

D.2 Results

As pre-registered, we screened participants for responding to temptation. We obtained 100 responses from participants who switch from complying (choosing the safe option) to violating (choosing the lottery) as temptation increases

[16]Participants in experiments expect to be asked meaningful questions. Therefore, asking to choose between receiving 0.05 GBP and incurring a certain loss may confuse participants into thinking that they have misunderstood the instructions, on top of increasing the number of questions by 50%. We, therefore, refrained from asking participants explicitly to make choices for $p_H = 1$.

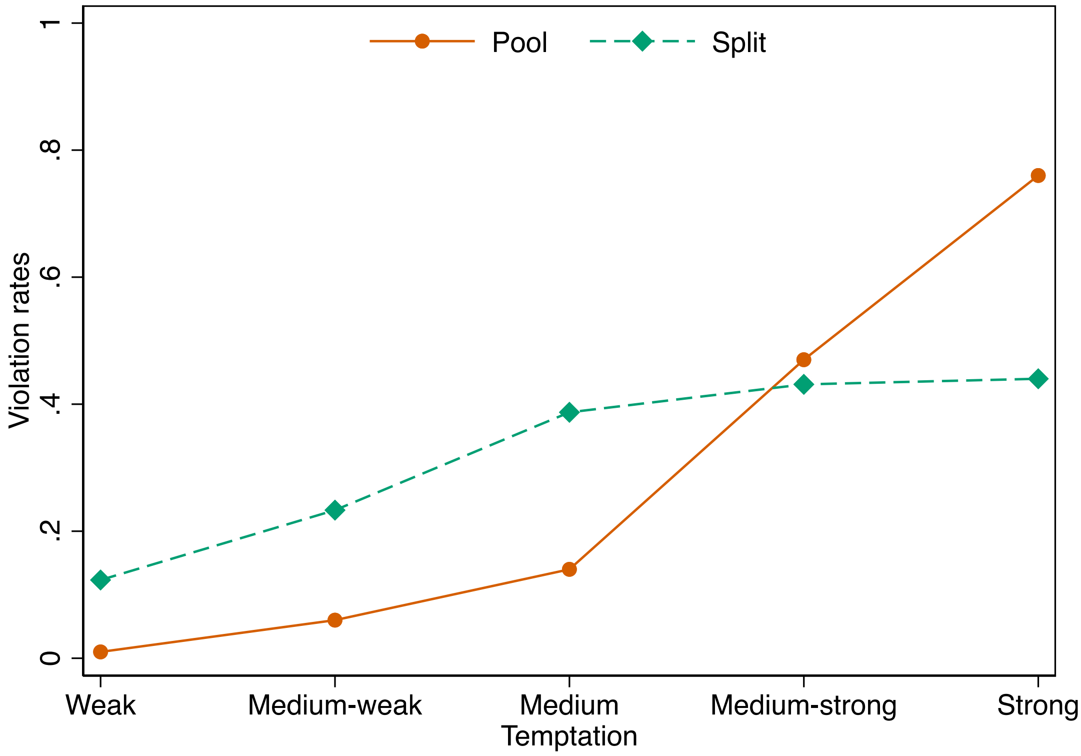


Figure D.1: Violation rates by temptation.

at least once, and never switched in the opposite direction. Note that screening was done only on the basis of temptation and not of splitting.

Aggregate violation rates. Recall that we set $\lambda = 0.44$. Since we assume full compliance under certain sanctioning, the expected violation rate under splitting is, therefore, zero if the participant choose to comply for the low sanctioning probability, and 0.44 if the participant chose to violate. Figure D.1 presents the average expected violation rates by splitting and temptation. We see that, in line with the theoretical predictions, temptation has a negligible effect on the violation rate under Medium-strong temptation, is socially beneficial for Strong temptation, and is socially detrimental for weaker levels of temptation.

Table D.1 reports the expected violation rates for pooling and for splitting with the difference due to splitting. The hypothesis testing reported is based

Table D.1: Violation rates under pooling and splitting.

	Pooling	Splitting	Difference	Robust Std. error
Weak	0.010	0.123	0.113^{***}	0.023
Medium-weak	0.060	0.233	0.173^{***}	0.032
Medium	0.140	0.387	0.247^{***}	0.036
Medium-strong	0.470	0.431	-0.039	0.051
Strong	0.760	0.440	-0.320^{***}	0.043

on an OLS regression of the choice to violate on a dummy for splitting interacted with dummies for temptation levels and robust standard errors clustered on individuals and post-estimation tests of the choice in the splitting question multiplied by 0.44 against the decision in the corresponding pooling question.[17] The results support the conclusion that splitting is socially detrimental for the lower levels of temptation and socially beneficial for Strong temptation.

Individual violation. At the individual level, splitting is socially beneficial if the participant chose to violate under the pooling sanctioning probability of 0.6; socially detrimental if the participant chose to violate for the low sanctioning probability but comply for the pooling probability; and socially benign if the participant always complies. Figure D.2 summarizes the distributions of these choice profiles by temptation level. Consistent with the theoretical prediction, the share of participants for whom splitting is socially beneficial increases with temptation. Splitting is detrimental for more individuals than it is beneficial for under Weak to Medium temptation and vice versa for Strong temptation. Under Medium-Strong temptation, splitting is beneficial and detrimtntal for approximately the same number of participants. Figure D.3 further presents the expected violation rates for each of the 100 individual participants.

[17] In our pre-registration we specified an equivalent analysis using a logistic regression. We report here the results from the linear regression to circumvent problems arising from the lack of variance in the answers to the Strong temptation with low probability of sanctioning, where all participants chose to violate. The results from the logistic regression are otherwise virtually identical, as are the results from a penalized maximum likelihood logistic regression with bootstrap standard errors clustered on individuals.

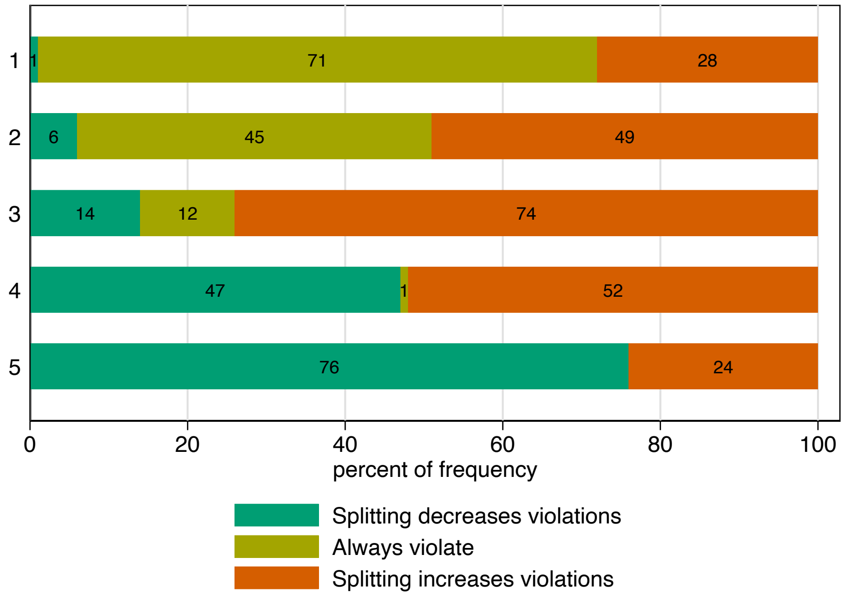


Figure D.2: Distribution of individual types.

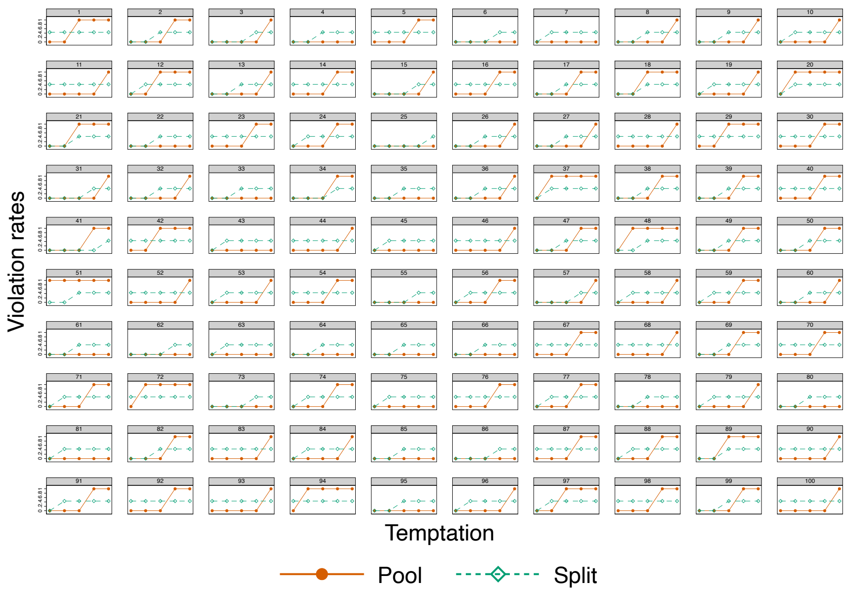


Figure D.3: Expected violations by individuals.