# Hannah Hörisch; Christina Strassmair:

# An experimental test of the deterrence hypothesis

Munich Discussion Paper No. 2008-4

Department of Economics
University of Munich

Volkswirtschaftliche Fakultät
Ludwig-Maximilians-Universität München

# An experimental test of the deterrence hypothesis*

Hannah Hörisch
University of Munich

Christina Strassmair
University of Munich†

February 26, 2008

**Abstract**

Crime has to be punished, but does punishment reduce crime? We conduct a neutrally framed laboratory experiment to test the deterrence hypothesis, namely that crime is weakly decreasing in deterrent incentives, i.e. severity and probability of punishment. In our experiment, subjects can steal from another participant's payoff. Deterrent incentives vary across and within sessions. The across subject analysis clearly rejects the deterrence hypothesis: except for very high levels of incentives, subjects steal more the stronger the incentives. We observe two types of subjects: selfish subjects who act according to the deterrence hypothesis and fair-minded subjects for whom deterrent incentives backfire.

Keywords: deterrence, law and economics, incentives, crowding out, experiment

JEL classification: K42, C91, D63

*We thank Francesco Drago, Dan Houser, Sandra Ludwig, Klaus Schmidt, Matthias Sutter, participants of the Theory Workshop at University of Munich, AFSE 2007 in Lyon, SMYE 2007 in Hamburg, MERSS 2007 in Mannheim, EDGE Jamboree 2007 in Cambridge, ENABLE YRM 2007 in Zurich for helpful comments on earlier versions of this paper. Financial support from SFB Transregio 15 and ENABLE *Marie Curie Research Training Network*, funded under the 6*th* Framework Program of the European Union, is gratefully acknowledged.

†*Affiliation:* University of Munich, Seminar for Economic Theory, Ludwigstraße 28 (Rgb.), 80539 Munich, Germany, email: hannah.hoerisch[at]lrz.uni-muenchen.de, christina.strassmair[at]lrz.uni-muenchen.de.

# 1 Introduction

That crime has to be punished seems to be universally accepted. The purpose and level of punishment, however, are controversial. Immanuel Kant advocated punishment to re-establish justice, Georg Friedrich Wilhelm Hegel stressed that ill has to be retaliated with ill. Both philosophers regard punishment as a mean to establish justice. In contrast, there exist schools of thoughts which stress that punishment shall prevent (future) crime. Becker's (1968) deterrence hypothesis is the classic economic contribution to the debate on punishment. According to Becker the purpose of punishment is to (efficiently) deter individuals from committing crimes. To achieve deterrence Becker relies on the power of pure deterrent incentives such as the severity and probability of punishment. The deterrence hypothesis states that crime rates fall in the severity and in the probability of punishment.

Our laboratory experiment tests the deterrence hypothesis in a controlled environment that permits to exogenously vary deterrent incentives, i.e. detection probability and level of punishment. For this purpose we use a very straightforward context, namely subjects have the possibility to steal from another subject's payoff. In this setup they cannot only decide whether they steal or not, but also how much they steal. We ask a very basic but important question: do deterrent incentives work?

In order to answer this question we have chosen one of the simplest possible designs: a modified dictator game. Two agents, $A$ and $B$, are randomly matched. Agent $A$ is a passive agent and has a higher initial endowment than agent $B$. Agent $B$ can decide how much to take away (steal) from $A$'s initial endowment. With probability $1 - p$, this amount is transferred from $A$ to $B$. With probability $p$ ("detection probability"), however, this amount is not transferred and a fixed fine $f$ is deducted from $B$'s initial endowment if $B$ has chosen a strictly positive amount.

We conduct six different treatments in which we vary detection probability $p$ and fine $f$. Our benchmark treatment T1 sets $p = f = 0$. Treatments T2, T3, and T4 investigate the range of small and intermediate deterrent incentives, i.e. levels of incentives such that taking agent $A$'s whole initial endowment pays off in expectation. Treatment T5 is characterized by a combination of $p$ and $f$ such that taking everything generates about the same expected payoff as taking nothing. In treatment T6, however, the expected payoff from taking everything is substantially smaller than the one from taking nothing. Each subject participates in two different treatments sequentially. This design permits both an across subjects and a within

subject analysis of taking behavior. In other words, we can analyze different regimes and regime changes with the data at hand.

Our experimental design focuses on the effects of simple and pure incentives which makes it distinct from previously conducted experiments. First of all, incentives are set exogenously (by the experimenter) and are not endogenously determined by another subject. Hence, the intensity of incentives does not signal trust, expectations or intentions which could potentially influence an agent's decision. Second, as one of the two agents is completely passive, any strategic uncertainty is removed for the active agent. Therefore, the intensity of incentives does not affect the active agent's beliefs about the other agent's choice and thereby affecting an agent's decision. Third, as the payoff table is common knowledge, the intensity of incentives does not signal costs and benefits of certain actions which could potentially drive an agent's decision. Our design has three main advantages. (i) The task is easy to understand for the subjects. (ii) Our design allows to test the isolated effect of incentives *per se*. (iii) It captures some crucial features of a certain class of crimes like stealing: the victim is rather passive, and it cannot set the severity of punishment and - to a large extent - the detection probability; in case of a theft the stolen amount is a good predictor of the thief's benefit and the victim's costs.

The results obtained in our across subjects analysis clearly reject the deterrence hypothesis: the average taken amount is not monotonically (weakly) decreasing in $p$ and $f$. In contrast, we find that incentives may backfire: on average subjects take significantly more in the treatment with intermediate deterrent incentives than in the absence of incentives. Only very strong incentives deter subjects from taking. The results of both our across and within subjects analysis can be explained by a model of two types: selfish subjects who react to deterrent incentives as predicted by the deterrence hypothesis and fair-minded subjects who take more when incentives are introduced or raised until incentives reach a very high level. Possible explanations for the behavior of the second type of subjects are crowding out of fairness concerns by extrinsic incentives or fairness concerns regarding expected outcomes. Only lasting crowding out of fairness concerns can explain the sequence effects in our data: many fair-minded subjects take more in a given treatment if this treatment was preceded by a treatment with stronger incentives than if it was preceded by a treatment with weaker incentives. Furthermore, we find that $p$ and $f$ seem to be interchangeable instruments in achieving deterrence.

Since we obtain our data from neutrally framed experiments, one may question our results and their applicability to "real life stealing". In real life crime and

deterrent incentives often have a strong moral connotation and policy makers may make use of that. Still, we consciously use a neutral frame because our primary aim is to test the economic approach to crime. Its core, the deterrence hypothesis, relies on pure incentive effects that are independent of all other factors that may influence crime. In Becker's (1968) model framing might ceteris paribus affect $B$'s decision, but not the comparative statics with respect to $p$ and $f$. Whatever the frame the taken amount should be monotonically decreasing in $p$ and $f$. In order to measure the effect of moral costs evoked by a non-neutral, "moral" framing, we run some additional framed sessions in which we label $B$'s decision as "stealing" if $x > 0$ and the fixed fine $f$ as "penalty" instead of "minus points". In these sessions, we still observe backfiring of incentives.

Becker's seminal paper has triggered numerous theoretical extensions as well as field studies testing its external validity.[1] At large the empirical literature implies that punishment reduces crime, but variations in detection probability and severity of punishment explain only a small part of the variation in crime (see Glaeser, 1999). This may be caused by methodological problems that arise when using field data. Usually only aggregate data are available which results in simultaneity bias[2] and omitted variable problems. Field data often report the behavior of offenders only and not that of the general population. Furthermore, measurement error is widespread as not all crime is reported. All these problems do not exist in the laboratory.

There already exist experimental studies focusing on criminal behavior. The experimental literature on tax evasion explicitly addresses deterrence.[3] The tax evasion setups clearly differ from ours though. In many settings subjects do not influence other subjects' payoffs at all, in other settings the collected taxes are used for public good provision or redistribution of resources among a group of subjects. In contrast, in our setup a stealing subject directly hurts another subject which seems to be a crucial feature of many crimes. Moreover, there are settings in which the tax authority is a player and can strategically interact with the taxpayer. In our setup, however, the victim is passive and incentives are set exogenously. Laboratory experiments on criminal behavior other than tax evasion are scarce. Falk and Fischbacher

[1] Garoupa (1997) and Polinsky and Shavell (2000a) provide comprehensive overviews on the economic theory of optimal law enforcement. Eide (2000) and Glaeser (1999) survey empirical studies of the deterrence hypothesis.

[2] See Levitt (1997) for a convincing example of how to address the simultaneity problem.

[3] Torgler (2002) reviews the experimental literature on tax evasion and concludes that evidence on the effectiveness of deterrent incentives is rather mixed (p.662).

(2002) explore the influence of social interaction phenomena on committing a crime. Bohnet and Cooter (2005), Tyran and Feld (2006), and Galbiati and Vertova (2005) investigate whether law can act as "expressive law", i.e. prevent crime by activating norms that prohibit committing a crime. Tyran and Feld (2006) also compare the effects of exogenously imposed and endogenously chosen incentives. Falk and Fischbacher (2002) and Bohnet and Cooter (2005) do not vary incentives and therefore can not test for incentive effects. The setups of Galbiati and Vertova (2005) and Tyran and Feld (2006) vary incentives, however in the context of a public good game which is much more complex than our setup.

In addition, there is a growing economic literature that investigates the effectiveness of incentives in different contexts as e.g. labor market relations. Some laboratory and field experiments document that (small) incentives backfire and thus challenge the belief in the effectiveness of incentives.[4] Frey and Jegen (2001) stress that introducing incentives has two countervailing effects: besides the standard relative price effect, incentives may crowd out intrinsic motivation. With small incentives the relative price effect is small and the latter, counterproductive effect may dominate. Since the contexts of those studies differ, the used setups are usually richer than ours: the incentives are often determined endogenously, an action's costs and benefits may not be common knowledge, or strategic interaction may cause strategic uncertainty.

The paper proceeds as follows. Section 2 presents the experimental design and procedure, section 3 the behavioral predictions and hypotheses. The across and within subjects analyses are summarized and discussed in section 4. In section 5 we check the robustness of our results by presenting results from sessions with a moral frame. Section 6 concludes.

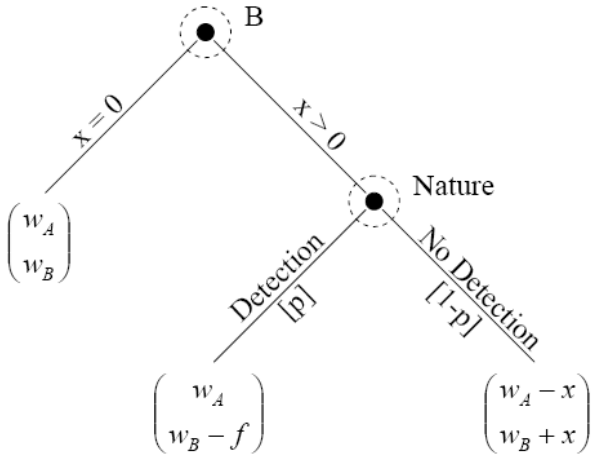
# 2 Experimental design and procedure

Consider the simplest possible stealing game with two agents, $A$ and $B$. Agent $A$ is initially endowed with $w_A$, and agent $B$ is initially endowed with $w_B$, where

[4] Bowles (2007), Fehr and Falk (2002), and Frey and Jegen (2001) survey the economic literature on crowding out of intrinsic motivation. The origins of this literature are in psychology, see for example Deci (1971) and Lepper et al. (1973). Deci et al. (1999) provide a meta-analysis of more than 100 psychological studies on the effect of extrinsic rewards on intrinsic motivation.

$w_A > w_B$.[5] While agent $A$ is passive, agent $B$ can take any amount $x \in [0, w_A]$ from agent $A$'s endowment. If $B$ does not take anything, i.e. $x = 0$, agents $A$ and $B$ both receive their initial endowments $w_A$ and $w_B$, respectively. If $B$ takes a strictly positive amount, i.e. $x > 0$, with probability $(1 - p) \in [0, 1]$ the taken amount $x$ is indeed transferred from $A$ to $B$; with probability $p$, however, the taken amount $x$ is not transferred and, on top of that, agent $B$ has to pay a fixed fine $f$. We use a fixed fine $f$ that is independent of $x > 0$ in order to keep the design as simple and clear as possible. The structure of the game is summarized in Figure 1.

Figure 1: Structure of the game




Since we focus on pure incentive effects on $B$'s behavior, we vary the detection probability $p$ and the fine $f$ across different treatments and fix $w_A$ and $w_B$ at levels 90 and 50, respectively. Table 1 presents the treatments.

Treatment T1, our benchmark treatment, implements no deterrent incentives. It is simply the mirror image of a dictator game.[6] In all other treatments a strictly positive $p$ and a strictly positive $f$ is implemented. We categorize the intensity of these incentives according to agent $B$'s expected payoff when taking agent $A$'s whole initial endowment. As one can see in Table 1, the level of incentives is (weakly) gradually increasing in the order of the treatment. In treatments T2, T3 and T4 taking everything pays off in expectation. Treatment T5 is characterized by a combination of $p$ and $f$ such that taking the maximally possible amount generates about the same expected payoff as taking nothing. In treatment T6, however, the expected payoff

[5] $w_A > w_B$ allows to distinguish between subjects who have a preference for fair (equal) outcomes and subjects who simply do not want to take anything in treatment T1.

[6] Here, subjects can decide how much to take away from (instead of to give to) another agent in a purely distributional context without any strategic considerations.

Table 1: Treatments

| Treatment | p | f | B's expected payoff given x = 0 | B's expected payoff given x = 90 | Level of incentives |
|---|---|---|---|---|---|
| T1 | 0.0 | 0 | 50 | 140 | zero |
| T2 | 0.6 | 6 | 50 | 82.4 | small |
| T3 | 0.5 | 25 | 50 | 82.5 | small |
| T4 | 0.6 | 20 | 50 | 74 | intermediate |
| T5 | 0.7 | 40 | 50 | 49 | high |
| T6 | 0.8 | 40 | 50 | 36 | very high |

from taking everything is substantially smaller than the one of taking nothing. Since the same intensity of incentives is implemented by different $p$ and $f$ in treatments T2 and T3, we can analyze whether $p$ and $f$ are interchangeable instruments in deterring taking behavior, at least for this level of incentives.

Each experimental session consisted of three parts: two different treatments of the stealing game and a dictator game.[7] After these three parts, participants filled out a questionnaire eliciting data on their age, sex and subject of studies. We used a paid Holt and Laury (2002) procedure to get an indication of subjects' risk preferences.[8] The conducted sessions are presented in Table 2.

At the beginning of each session, participants were told that one randomly picked part out of the three would be paid for all of them. After each part, only the instructions for the following part were handed out. Subjects did not receive any feedback before the end of the experiment. They were matched according to a perfect stranger design, i.e. a couple matched once is never matched again in the following parts. Those subjects randomly chosen to be agents $B$ in part 1 remained agents $B$ in part 2 and were assigned the role of the dictator in part 3. Consequently, passive subjects remained passive throughout all three parts of the session.[9]

This design offers the possibility to analyze the observed behavior in two different ways. First, we can compare behavior in part 1 across different treatments. This

[7] In the standard dictator game, the dictator could give any amount of his initial endowment of 90 to a passive agent with an initial endowment of 50. The chosen amount may indicate the dictator's aversion to advantageous inequity. However, the donated amount might be affected by the treatments played in part 1 and part 2.

[8] The translated table and a brief report on the observed levels of risk aversion can be found in the appendix.

[9] To keep the passive subjects busy we asked them how they would decide if they were agent $B$.

Table 2: Session plan

| Session | Part 1 | Part 2 | Part 3 | Questionnaire* | Number of participants |
|---|---|---|---|---|---|
| T1T3 | T1 | T3 | DG | Yes | 38 |
| T3T1 | T3 | T1 | DG | Yes | 38 |
| T2T3 | T2 | T3 | DG | Yes | 18 |
| T3T2 | T3 | T2 | DG | Yes | 20 |
| T2T4 | T2 | T4 | DG | Yes | 38 |
| T4T2 | T4 | T2 | DG | Yes | 36 |
| T5T6 | T5 | T6 | DG | Yes | 32 |
| T6T5 | T6 | T5 | DG | Yes | 38 |

* includes a Holt and Laury (2002) table

is the cleanest comparison because individual behavior in part 1 is not influenced by any preplay. Second, we can analyze how agents $B$ adapt their behavior to the change in incentives from part 1 to part 2. Since the structure of the game is very simple, we assume that a change in behavior from part 1 to part 2 is stimulated by the change of incentives rather than learning.

Our experimental sessions were run in November 2006 and March 2007 at the experimental laboratory of the SFB 504 in Mannheim, Germany. 258 students of the Universities of Mannheim and Heidelberg participated in the experiment. Subjects were randomly assigned to sessions and could take part only once. The sessions were framed neutrally[10] and lasted about 40 minutes. Subjects did not receive a show-up fee[11] and earned 12.34 € on average.

# 3 Behavioral predictions and hypotheses

We focus on the question how the intensity of incentives affects $B$'s decision. This depends on the specific form of $B$'s utility function. Different theoretical approaches make different assumptions concerning this point and, therefore, have varying behavioral predictions.

[10] Translated instructions for players $B$ can be found in the appendix.

[11] Six subjects did not earn anything in the randomly selected part and in the Holt and Laury (2002) table.

## 3.1 Behavioral predictions

**Model 1: The self-interest model**

The standard neoclassical approach assumes that all people are selfish, i.e. their utility function $U$ depends on their own material payoff $m$ only and is increasing in $m$.

With these assumptions the deterrence hypothesis holds, namely the optimal taken amount $x^*(p, f)$ is monotonically (weakly) decreasing in $p$ and in $f$.

Due to the fixed fine $f$ agent $B$ who maximizes his expected utility either takes as much as possible ($w_A$) or nothing. This depends on the relative sizes of $p$, $f$, $w_A$, $w_B$ and on the level of risk aversion. $B$'s optimal taken amount is

$$x^*(p,f) \in \begin{cases} \{0\} & if \quad p > \frac{U(w_A+w_B)-U(w_B)}{U(w_A+w_B)-U(w_B-f)} \\ \{0, w_A\} & if \quad p = \frac{U(w_A+w_B)-U(w_B)}{U(w_A+w_B)-U(w_B-f)} \\ \{w_A\} & if \quad p < \frac{U(w_A+w_B)-U(w_B)}{U(w_A+w_B)-U(w_B-f)} \end{cases} .$$

The higher $p$ and/or the higher $f$, the less attractive it is to take everything. For sufficiently high values of $p$ and $f$, agent $B$ does not take anything. This holds for any risk preferences, i.e. it is independent whether $U$ is concave or convex in $m$. A higher level of risk aversion[12] reduces $\frac{U(w_A+w_B)-U(w_B)}{U(w_A+w_B)-U(w_B-f)}$ ceteris paribus, and thus the set of $p$, $f$, $w_A$, $w_B$ combinations for which taking everything is optimal.

Empirical studies have shown that individual behavior may systematically deviate from predictions of the standard neoclassical approach. In these contexts, however, observed behavior may be consistent with predictions of models of other-regarding preferences.[13] Our two-agent setup with unequal initial endowments is one of the contexts in which it is very plausible to consider models of fairness concerns.

**Model 2: A model of fairness concerns regarding final outcomes**

Models of fairness concerns regarding final outcomes (e.g. Fehr and Schmidt, 1999, Bolton and Ockenfels, 2000) assume that an agent's utility function $\widetilde{U}$ is increasing in the agent's own material payoff $m$, but decreasing in the material payoff inequality $|m - y|$ with $y$ as the material payoff of the other agent.

The deterrence hypothesis still holds for a very general class of these models, i.e. if there exists a unique optimal decision $x^*(p = 0, f = 0)$ maximizing agent $B$'s expected utility for $p = 0$ and $f = 0$.

[12] Consider the Arrow-Pratt measure of relative risk aversion $-\frac{m*U''(m)}{U'(m)}$, for example.

[13] Fehr and Schmidt (2006) survey empirical foundations of other-regarding preferences.

Due to the fixed fine $f$ agent $B$ who maximizes his expected utility either takes an amount which is optimal for no incentives ($x^*(p = 0, f = 0)$) or nothing. For relatively low values of $p$ and $f$, agent $B$ takes $x^*(p = 0, f = 0)$ that may be smaller than $w_A$. For relatively strong incentives, agent $B$ is deterred and takes nothing.

The reason is that agents cannot trade off payoffs from different states, in our context payoffs if $B$'s taking is detected and payoffs if $B$'s taking is not detected. Then, $x^*(p \geq 0, f \geq 0)$ cannot be strictly larger than $x^*(p = 0, f = 0)$: if $B$'s taking is not detected, $\widetilde{U}$ is maximized at $x^*(p = 0, f = 0)$; if $B$'s taking is detected, utility $\widetilde{U}$ is the same for any $x > 0$ and larger for $x = 0$. Analogously, taking any strictly positive, but strictly smaller amount than $x^*(p = 0, f = 0)$ yields less expected utility than taking $x^*(p = 0, f = 0)$ and, therefore, cannot be optimal.

**Model 3: A model of fairness concerns regarding expected outcomes**

Models of fairness concerns regarding expected outcomes (e.g. Trautmann 2007) assume that an agent's utility function $\widehat{U}$ is increasing in the agent's own material payoff $m$ and decreasing in the absolute difference between own expected payoff $m^e$ and the other agent's expected payoff $y^e$ ($|m^e - y^e|$).[14]

If $|m^e - y^e|$ directly enters the utility function, the deterrence hypothesis may not hold any more, i.e. there may exist a value of $p$ and $f$ for which $x^*(p, f)$ is *strictly* increasing in $p$ and/or in $f$.

The reason is that agents can trade off payoffs from different states, e.g. an advantageous inequity in material payoffs if $B$'s taking is not detected can compensate a disadvantageous inequity in material payoffs if $B$'s taking is detected. As an illustration consider the following utility function $\widehat{U} = m - \beta * max\{m^e - y^e, 0\}$ with $m^e = (1 - p) * (w_B + x) + p * (w_B - f)$ and $y^e = (1 - p) * (w_A - x) + p * w_A$. If $\beta > \frac{1}{2}$, agent $B$ who maximizes his expected utility tries to perfectly equate $m^e$ and $y^e$ by choosing $x$. Hence, agent $B$ takes more, the higher $p$ and/or the higher $f$. Nevertheless, deterrence by strong incentives may still occur in this illustration as $x$ is bounded above by $w_A$.[15]

**Model 4: A model of fairness concerns (regarding final outcomes) that are crowded out by incentives**

The literature on crowding out of intrinsic motivation by extrinsic incentives uses the term "intrinsic motivation" very broadly. It may apply to fairness concerns

[14] Therefore, the evaluation of a state may not be independent from another state.

[15] Consider $\widehat{U}$ and assume that $p$ and $f$ are so high that taking everything would yield $m^e < y^e$ with $m^e < w_B$. In this case taking nothing is optimal.

as well. In our context, crowding out implies that agents' fairness concerns are diminishing in the intensity of deterrent incentives. Formally, this assumption can be captured by the following utility function:

$$V = \lambda(p, f) * U(m) + [1 - \lambda(p, f)] * \widetilde{U}(m, |m - y|),$$

where as above $U(m)$ represents utility of a selfish agent and $\widetilde{U}(m, |m - y|)$ utility of an agent with fairness concerns regarding final outcomes. The core of the crowding out assumption is that $\lambda(p, f) \in [0, 1]$, the weight of $U(m)$, is increasing in $p$ and in $f$.

With these assumptions, there may be ranges of $p$, $f$ combinations such that the optimal amount $x^*(p, f)$ *strictly* increases in $p$ and/or in $f$. Therefore, the deterrence hypothesis does not necessarily hold.

The intuition is that if $p$ and $f$ are relatively low, agent $B$ may be strongly affected by fairness concerns and take an interior amount of his choice set; if $p$ and $f$ increase, agent $B$ may be less affected by fairness concerns and may take a substantially higher amount. If the level of incentives is very high, though, such that both selfish and fair-minded subjects are deterred agent $B$ does not take anything.

Furthermore, the literature on crowding out of intrinsic motivation by extrinsic incentives has drawn attention to the following two aspects on which the verdict is still out.

**(i) Continuity of crowding out**

$\lambda(p, f) \in [0, 1]$ may increase continuously in $p$ and in $f$. Even if this is the case $x^*(p, f)$ may increase discontinuously in $p$ and in $f$ for some $\widetilde{U}(m, |m - y|)$.

The empirical results of Gneezy and Rustichini (2000b) and Gneezy (2003) suggest discontinuous crowding out. Frey and Oberholzer-Gee (1997), however, explain their data by assuming continuous crowding out.

In our context, subjects who increase their taken amount $x$ to a level strictly less than the maximal amount of $w_A$ as a reaction to an introduction or an increase in incentives are evidence for continuous rather than discontinuous crowding out.

**(ii) Hysteresis**

Extrinsic incentives may crowd out fairness concerns lastingly. As a consequence the crowding out effect of an increase in incentives is larger than the crowding in effect of the *subsequent* decrease in incentives that reverses the increase in incentives by size.

Some studies (e.g. Irlenbusch and Sliwka, 2005, Gneezy and Rustichini, 2000, Gächter, Königstein, and Kessler, 2005) find evidence for hysteresis, i.e. evidence

that incentives crowd out fairness concerns lastingly.

In our context, if subjects take "fairer" amounts in a given treatment played in part 1 than subjects in the same treatment played in part 2 after a part 1 with relatively stronger incentives, this is evidence for hysteresis. Note that backfiring of incentives *and* hysteresis can only be explained by a model of lasting crowding out of fairness concerns and not by a model of fairness concerns regarding expected outcomes. Thus, hysteresis might be a mean to distinguish between these two models (models 3 and 4) that can explain backfiring of incentives.

## 3.2 Hypotheses

The predictions of the various models differ. However, all four models predict hypothesis 1.

**Hypothesis 1: Deterrence by strong incentives**

Relatively high values of the detection probability $p$ and the fixed fine $f$ deter agents from taking. This range is larger, the more risk averse an agent is.

The threshold of strong incentives may vary by subject. A risk neutral or risk averse selfish agent abstains from taking in treatments T5 and T6. A risk neutral or risk averse agent with standard Fehr-Schmidt (1999) fairness preferences may even abstain from taking in treatment T4. A subject with fairness concerns regarding expected outcomes may only abstain from taking if $p = 1$ and $f > 0$.

In contrast to hypothesis 1, hypothesis 2 is necessarily implied by the self-interest model and the model of fairness concerns regarding final outcomes, but not by the model of fairness concerns regarding expected outcomes and the model of fairness concerns that are crowded out by incentives.

**Hypothesis 2: Deterrence hypothesis**

The taken amount $x$ is monotonically (weakly) decreasing in the detection probability $p$ and the fixed fine $f$.

Hypothesis 2 implies that the average taken amount $x$ should be (weakly) decreasing from treatments T1 to T6.

# 4 Results

In a first step we compare behavior in part 1 across subjects. This step has the advantage that behavior is not influenced by any preplay. However, we cannot draw any conclusion whether the observed phenomena are valid on an individual basis, and whether hysteresis occurs. In a second step we address these issues by comparing behavior in part 1 with behavior in part 2.

## 4.1 Comparison of treatments in part 1

### 4.1.1 Summary statistics

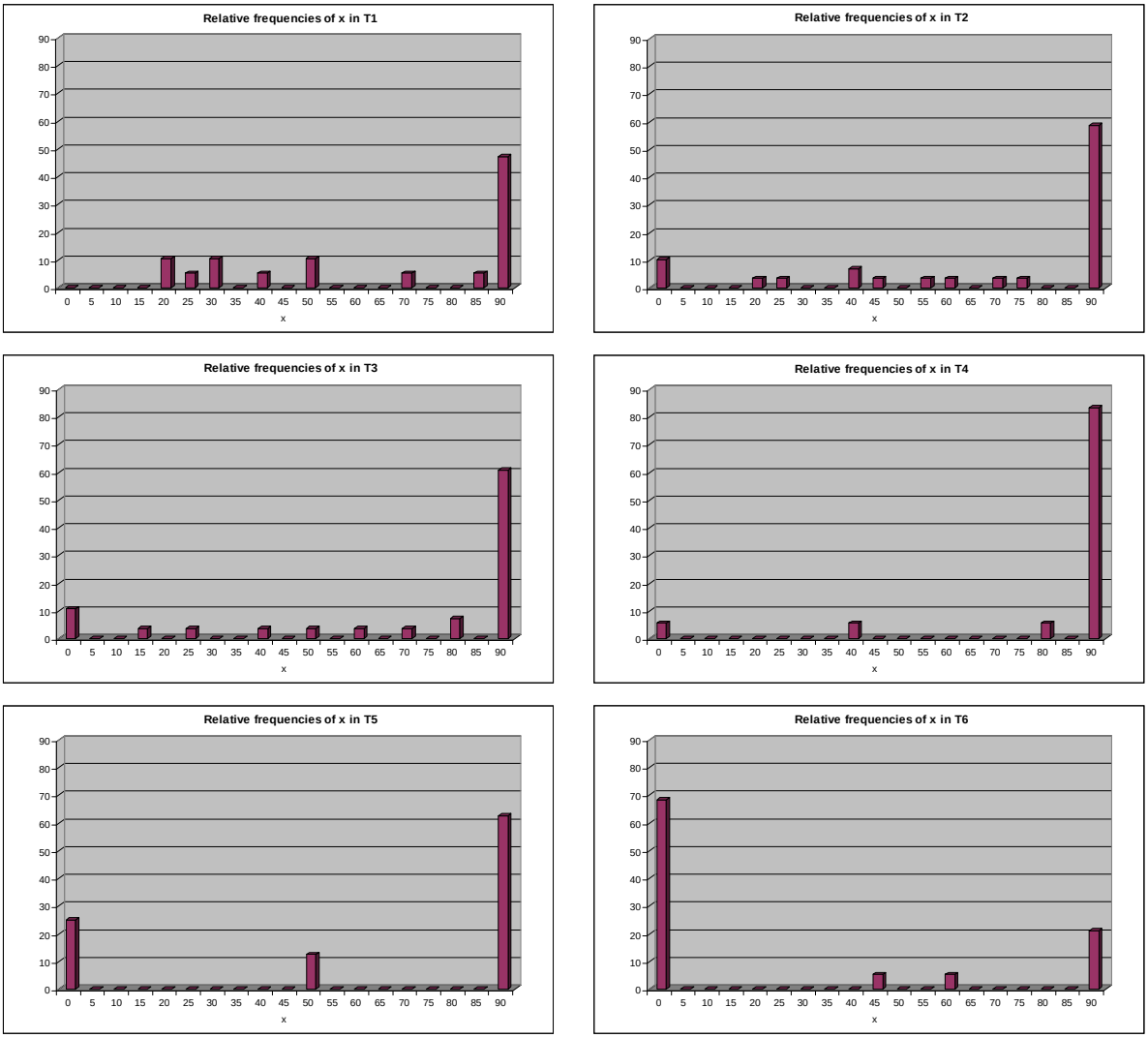
**Benchmark treatment**

The experimental data in treatment T1 show how much people take in the absence of deterrent incentives. The upper left panel of Figure 2 summarizes the distribution of the taken amount $x$ in the benchmark treatment.

As treatment T1 is the mirror image of a dictator game, we can compare behavior in T1 with standard results of dictator games as for example Forsythe et al. (1994). In line with their paper, we can identify two types of agents: selfish agents and fair-minded agents. In their benchmark treatment (the paid dictator game conducted in April with a pie of 5 $) about 45 % of subjects are "pure gamesmen" who do not give anything, and the rest gives a strictly positive amount. These types of agents correspond remarkably well to the 47 % (52.5 %) of selfish subjects in treatment T1 who take everything (between 80 and 90), and the rest who takes a strictly positive amount below 90 (80).

To summarize, we have two types of agents: slightly less than 50 % of our subjects have selfish preferences while a bit more than 50 % have fairness concerns. As the model of fairness concerns regarding final outcomes shows fairness concerns are not necessarily a reason why the deterrence hypothesis might fail. But it may fail if fairness concerns are based on expected outcomes or if they are crowded out by deterrent incentives. To figure out whether this is the case we have a closer look at the treatments with deterrent incentives.

Figure 2: Distributions of $x$ per treatment (in intervals of size 5)[16]




**Treatments with deterrent incentives**

Figure 3 summarizes the average taken amount $x$ per treatment. Treatments are ordered by the intensity of deterrent incentives, i.e. the combined effect of detection probability $p$ and fine $f$ (compare Table 1). The average taken amount $x$ increases in the range of no, small, and intermediate incentives (from T1 to T4), while it decreases in the range of strong and very strong incentives (T5 and T6). Hence, the relationship between the average taken amount and the intensity of deterrent incentives is rather inverted-U shaped than monotonically decreasing.

[16]Interval $y < 90$ denoted on the horizontal axis is the union of all points $x \in [y, y + 5)$. Interval $y = 90$, in contrast, is the union of all points $x = 90$.

Figure 3: Average taken amount $x$ per treatment

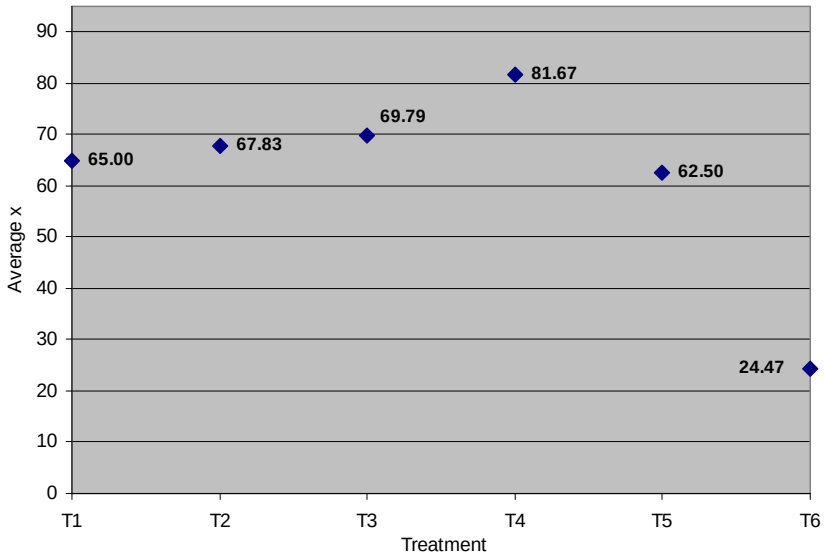



Figure 2 shows that the fraction of subjects taking everything increases by treatment from T1 to T4. In treatment T4, it peaks at a value of more than 80 % which is considerably higher than the corresponding 47 % in the absence of any incentives as in treatment T1. From treatment T5 onwards, this fraction decreases.

Still, the share of subjects not taking anything monotonically increases in the level of incentives. It is moderate with no, small and intermediate incentives ($\leq$ 10 %), quite substantial with strong incentives (about 25 %), and largest with very strong incentives (nearly 70 %).

Interestingly, there are always subjects taking interior values of their strategy set, most so in the benchmark treatment. The share of these subjects decreases in the intensity of incentives. Moreover, the average taken amount conditional on interior values increases by treatment from T1 to T4.

Compared to the benchmark treatment deterrent incentives shift mass to the borders of the support. We observe both backfiring of small incentives and deterrence at the same time.[17] Small and intermediate incentives move mass predominately towards the upper border which stands in sharp contrast to the deterrence hypothesis, but is consistent with models 3 and 4. Strong and very strong incentives move mass exactly to the lower border which is consistent with hypothesis 1.

Since the results of treatments T2 and T3 are very similar, detection probability and fine seem to be interchangeable instruments.

[17] In an experiment on corruption that uses probabilistic incentives as we do, Schulze and Frank (2003) observe a similar pattern in their data.

### 4.1.2 Analysis of hypotheses

A Kruskal-Wallis test on behavior in part 1 documents significant ($p < 0.01$) treatment effects. In order to identify and characterize the significant differences we run pairwise Mann-Whitney-U tests. The one-sided p-values are recorded in Table 3.

Table 3: One-sided p-values of pairwise Mann-Whitney-U tests

| | T2 | T3 | T4 | T5 | T6 |
|---|---|---|---|---|---|
| T1 | 0.287 | 0.234 | 0.015 | 0.400 | $< 0.001$ |
| T2 | | 0.408 | 0.040 | 0.447 | $< 0.001$ |
| T3 | | | 0.058 | 0.390 | $< 0.001$ |
| T4 | | | | 0.071 | $< 0.001$ |
| T5 | | | | | 0.005 |

In treatment T6, agents take significantly ($p < 0.01$) less than in any other treatment. This is consistent with hypotheses 1 and 2. However, contradictory to hypothesis 2, the deterrence hypothesis, agents take significantly more in treatment T4 than in treatments T1 ($p < 0.05$), T2 ($p < 0.05$) and T3 ($p = 0.058$).[18] There is no significant difference in behavior in treatments T2 and T3.

In order to account for individual characteristics when comparing treatments we estimate two specifications whose results are presented in Table 4.

First, we regress the taken amount $x$ on individual characteristics and treatment dummies using an OLS estimation with robust standard errors. Second, we address the fact that the taken amount $x$ is truncated and estimate a Tobit specification with the same regressors.

In both estimations the treatment dummy for T4 is significantly positive ($p < 0.05$), the treatment dummy for T6 is significantly negative ($p < 0.05$), and the treatment dummies for T2 and T3 are not significantly different from each other.[19] Hence, these results are robust. Risk aversion has a significantly negative effect ($p < 0.05$) on the taken amount in both specifications (as subjects with a high level of

[18] One-sided Kolmogorov-Smirnov tests and $\chi^2$-tests based on a grouping of subjects according to whether they are deterred, try to roughly equate payoffs (take between 15 and 29 units), show some fairness concerns (take between 30 and 79 units) or are selfish (take between 80 and 90 units) largely confirm the results of the Mann-Whitney-U tests presented here. In particular, subjects always take significantly more in treatment T4 than in T1.

[19] The inclusion of interaction effects of the dummy for risk aversion with the treatment dummies in the OLS estimation with robust standard errors does not change any of these results.

Table 4: Regression results

| Dependent variable: x | OLS-r | Tobit |
|---|---|---|
| Intercept | +057.03*** | +094.54*** |
| Sex (1 if male, 0 else) | +012.14* | +028.08 |
| Risk aversion (1 if risk averse, 0 else) | - 014.55** | - 057.16** |
| Economist (1 if economist, 0 else) | +010.05* | +030.48 |
| DG (donated amount in part 3) | - 000.12 | - 000.48 |
| T2 | +010.23 | +033.81 |
| T3 | +009.04 | +027.32 |
| T4 | +018.38** | +087.65** |
| T5 | - 007.38 | - 020.79 |
| T6 | - 042.74*** | - 132.64*** |
| Number of observations | 129 | 129 |
| (Pseudo) Rš | 0.3049 | 0.0754 |

Ti: 1 if treatment = Ti, 0 else
*, **, *** significant at 10, 5, 1 percent significance level
-r with robust standard errors

risk aversion are more likely to be deterred).

Given the results of the Mann-Whitney-U tests and the regressions we do not reject hypotheses 1, but we reject hypothesis 2, the deterrence hypothesis.

**Result 1: Deterrence by strong incentives**

Very strong incentives as in treatment T6 significantly reduce the taken amount. On average, risk averse agents take significantly less.

**Result 2: Backfiring of small incentives**

Deterrent incentives do not monotonically (weakly) decrease the average taken amount. Intermediate incentives in treatment T4 significantly increase the average taken amount.

**Result 3: Interchangeability of detection probability and fine**

We do not find any significant differences between treatments T2 and T3. In that sense, detection probability $p$ and fine $f$ seem to be interchangeable policy instruments.

In sum, these results are consistent with the predictions of the models 3 and 4.

## 4.2 Comparison of behavior in part 1 with behavior in part 2

Up to now we have compared different treatments across *different subjects* in part 1. In contrast to the deterrence hypothesis our results so far show that small and intermediate incentives backfire. Crowding out of fairness concerns or a model of fairness concerns regarding expected outcomes are explanations for this phenomenon. Since each subject sequentially participated in two different treatments, we can further analyze how *the same individuals* react to a change of deterrent incentives.[20] Sessions in which we increase incentives allow us to analyze (i) whether backfiring of small and intermediate incentives is observed on an individual level and (ii) whether backfiring is a continuous or discontinuous process. Sessions in which we decrease incentives enable us to check whether we observe hysteresis. Hysteresis can be explained by lasting crowding out of fairness concerns, but is inconsistent with the model of fairness concerns regarding expected outcomes. Sessions with incentives of the same intensity in both parts indicate whether $p$ and $f$ are interchangeable instruments on an individual level.

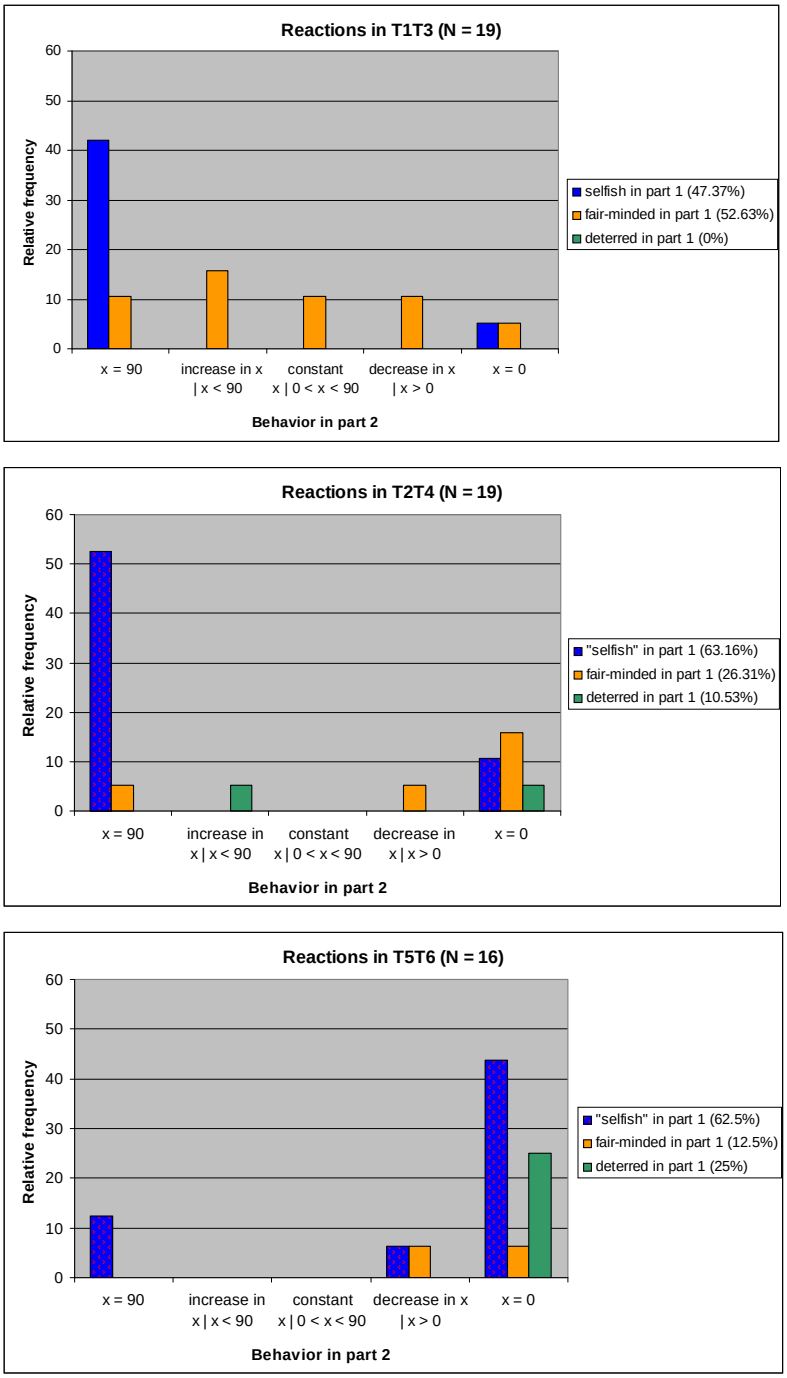
### 4.2.1 Backfiring of incentives on an individual level

In three different sessions, we increase incentives from part 1 to part 2: in session T1T3 from no to small incentives, in session T2T4 from small to intermediate incentives, in session T5T6 from strong to very strong incentives. Figure 4 summarizes how subjects behave in part 2 conditional on whether they acted selfishly ($x = 90$), acted fair-mindedly ($0 < x < 90$) or were deterred ($x = 0$) in part 1.

Since the benchmark treatment was played in the first part of session T1T3, we can identify about 47 % of subjects with selfish preferences. All except one take everything in part 2 again. 53 % of all subjects take an amount strictly less than everything in part 1. About a third of them increase the taken amount $x$ to a level smaller than 90, a fifth switches to taking everything in part 2, and another fifth keeps $x$ constant. Hence, for 50 % of fair-minded subjects small incentives seem to strictly backfire. Only one selfish and one fair-minded subject are deterred by small incentives.

[20] Since subjects do not get any feedback after part 1, behavioral effects cannot be triggered by the realization of punishment.

Figure 4: Reactions to an increase in the intensity of incentives




In session T2T4, about 63 % of subjects take everything already in part 1. We cannot distinguish whether they have selfish preferences or whether they have fairness concerns which are completely crowded out by the small incentives present in part 1. Again, the majority of these subjects is not deterred and keeps taking everything in part 2. The share of subjects taking intermediate amounts in part 1 is considerably smaller than in session T1T3. For 20 % of these subjects the increase of incentives completely backfires. The majority, however, is deterred. Note that a

moderate fraction of deterrence can already be found in part 1.[21]

In session T5T6, 62.5 % of subjects still take everything in part 1. More than two thirds of them are deterred by the increase of incentives though. 25 % of all subjects are deterred in part 1 and stay deterred in part 2. Only 12.5 % of subjects take a strictly positive amount below 90 in part 1. Half of them are deterred in the second part.

These observations can be summarized in the following two results:

**Result 4: Backfiring of small incentives on an individual level**

Subjects seem to be heterogeneous. There are selfish agents for which the deterrence hypothesis holds. However, there are also fair-minded agents for which small and intermediate incentives backfire. Independent of the type of agent, strong incentives deter.

**Result 5: Continuous and discontinuous backfiring of incentives**

We find evidence for both continuous and discontinuous backfiring of incentives.

### 4.2.2 Hysteresis

Whether hysteresis (lasting crowding out of fairness concerns) is present in our data can be seen by comparing behavior of a given treatment played in part 1 with behavior of the same treatment played in part 2 after a part 1 with stronger incentives. Hysteresis implies that we observe sequence effects for these treatments. Table 5 records two-sided p-values of pairwise Mann-Whitney-U tests that compare the *same* treatment played in *different* parts of a session.[22]

As Table 5 indicates we observe sequence effects in treatments T1, T2, and T5. Subjects in T1 take significantly ($p < 0.05$) more when it is played after T3 (81.3 instead of 65.0 units on average). Preplay in T3 with small incentives increases the average taken amount in treatment T1 that does not implement any incentives. Similarly, the average taken amount in T2 is significantly ($p < 0.05$) higher when it is played second (after a harsher or a constant intensity of incentives) than first. Both

[21] In sessions T1T3 and T2T4, none of the proposed models can explain the behavior of subjects who react to increased incentives by decreasing the taken amount to a level strictly larger than 0 or increasing it from 0 to a strictly positive amount.

[22] Treatments T2 and T3 are played second in two different sessions. Since the observations from the different second parts are not significantly different (p=0.71 and p=0.34, respectively according to two-sided Mann-Whitney-U tests) for different sessions, we do not report each session comparison separately.

Table 5: Non-parametric comparisons of different sequences (Mann-Whitney-U test)

| Treatment | played first in | played second in | p-value (two sided) |
|---|---|---|---|
| T1 | T1T3 | T3T1 | 0.082 |
| T2 | T2T3<br>T2T4 | T3T2<br>T4T2 | 0.099 |
| T3 | T3T1<br>T3T2 | T1T3<br>T2T3 | 0.676 |
| T4 | T4T2 | T2T4 | 0.061 |
| T5 | T5T6 | T6T5 | 0.014 |
| T6 | T6T5 | T5T6 | 0.617 |

results are consistent with a model of lasting crowding out of fairness concerns, but cannot be reconciled with the predictions of a model of fairness concerns regarding expected outcomes. In contrast, in T5 with strong deterrent incentives subjects take significantly ($p < 0.05$) less when it is played after T6. Preplay in T6 with very strong incentives seems to increase deterrence in treatment T5. This is inconsistent with a model of fairness concerns regarding expected outcomes and is inconsistent with lasting crowding out of fairness concerns if fairness concerns imply less taking in treatment T5 than selfish concerns imply.

**Result 6: Hysteresis**

Small and intermediate incentives have a lasting effect. They still backfire when incentives are decreased or even removed in the following period.

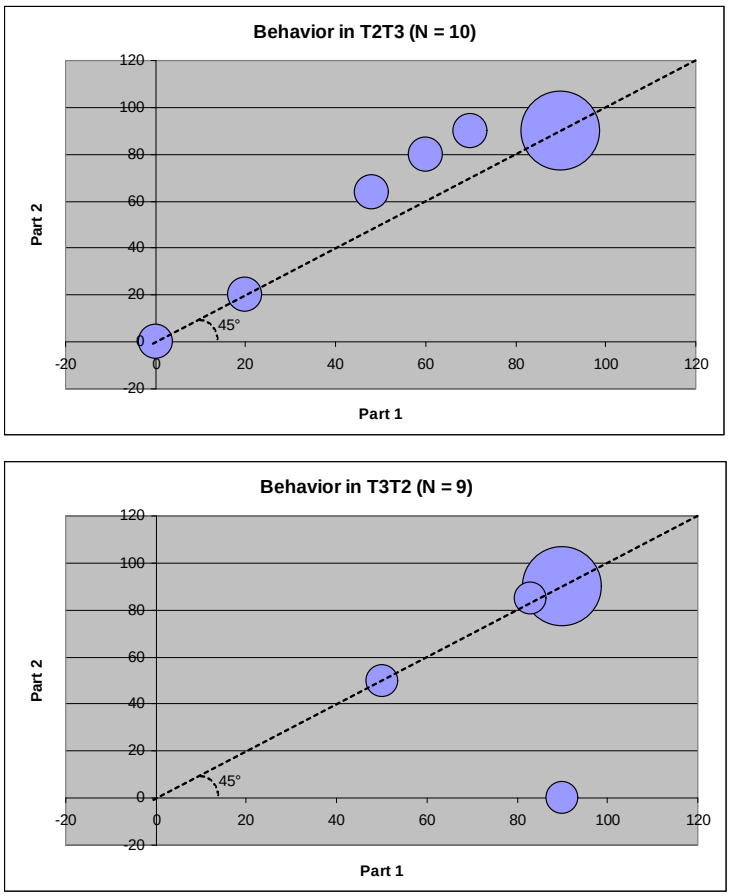
Since we observe hysteresis, a model of fairness concerns regarding final outcomes that is crowded out by incentives explains our data better than a model of fairness concerns regarding expected outcomes. Hysteresis also underlines how costly extrinsic incentives are. In addition to the effect incentives have in the current period they may also influence behavior in future periods. From this perspective, also strong and very strong incentives could potentially backfire by crowding out fairness concerns in future periods in which incentives are smaller.

In treatments with an increase in incentives there are no significant sequence effects for treatments T3 and T6, but subjects in treatment T4 take significantly ($p < 0.05$) less when it is played in part 2 after treatment T2 than when treatment T4 is played in part 1.

### 4.2.3 Substitutability of detection probability and fine

Since treatments T2 and T3 have the same intensity of deterrent incentives implemented by different values of the detection probability $p$ and fine $f$, we can test - at least for this specific level of incentives - whether these two instruments are interchangeable. We have already observed that the treatments T2 and T3 do not differ significantly across subjects in part 1 (result 3). Our within subject analysis in Figure 5 confirms this result.

Figure 5: Reactions to a change in incentives keeping their intensity constant




In session T2T3, 7 out of 10 subjects do not change their behavior. In session T3T2, only a single subject is apart from the 45° line. 6 subjects keep taking everything, 2 keep taking the same intermediate amount.

**Result 7: Interchangeability of detection probability and fine on an individual level**

Our within subjects comparison confirm result 3 that $p$ and $f$ seem to be interchangeable instruments.

# 5 Robustness check - Framing

So far we have presented results from neutrally framed experiments. This is a valid approach to test the deterrence hypothesis that relies on pure incentive effects that are independent of all other factors that may influence crime as e.g. the frame. While a non-neutral frame may ceteris paribus affect the taken amount (e.g. due to additional moral costs), comparative statics should remain unchanged. For any given (neutral or non-neutral) frame the deterrence hypothesis predicts the taken amount to be monotonically decreasing in detection probability and fine. In contrast, it is not clear whether a non-neutral frame interacts with incentives in the model of fairness concerns regarding final outcomes that are crowded out by incentives which fits our data best. While neutrally framed incentives crowd out fairness concerns, this may not necessarily be the case for incentives that are combined with a strong moral connotation.

In real life deterrent incentives often have a moral connotation and policy makers may try to make use of that. This is why we run two additional morally framed sessions and have a look at whether a non-neutral, moral frame will change our results. In the morally framed sessions $B$'s decision was labeled as "stealing" if $x > 0$ and the fixed fine $f$ was called "penalty" instead of "minus points". Apart from these two different labels the neutrally and morally framed sessions were conducted completely identically. In order to check whether framing affects behavior in the absence of incentives we run a framed version of treatment T1 (T1f). To analyze whether framing and incentives interact we run a framed version of treatment T4 (T4f).[23] 38 subjects participated in session T1fT4f, 32 subjects in session T4fT1f.

The results in the framed and neutral treatments are similar. There is no significant framing effect in part 1 in the absence of incentives, i.e. between T1 and T1f (two-sided Mann-Whitney-U test: $p > 0.5$). In contrast, subjects take more in part 1 in treatment T4 than in treatment T4f (two-sided Mann-Whitney-U test: $p = 0.075$). There is no significant difference in parts 1 between treatments T1f and T4f. However, the within subjects analysis documents a substantial degree of crowding out: when incentives are introduced in part 2 of session T1fT4f more than 30 % of individuals flip from taking intermediate amounts to taking everything. This parallels the results obtained in the neutrally framed sessions T1T3 and T2T4. In sum,

[23] We choose treatment T4 since the intensity of deterrent incentives in this treatment is (i) low enough not to deter the majority of individuals and (ii) high enough to potentially crowd out fairness concerns significantly.

we conclude that also with moral framing backfiring of intermediate incentives is a non-negligible phenomenon.

# 6 Conclusion

We have presented an experimental test of the deterrence hypothesis applied to the context of stealing. Our across subjects analysis of part 1 reject the hypothesis that the average taken amount is monotonically decreasing in deterrent incentives. On average, subjects take most when intermediate incentives are present. Only very strong incentives deter.

Both our across subjects comparison of behavior in part 1 and our within subjects comparison of behavior in part 1 with behavior in part 2 reflect two different types of subjects. We identify 50 % selfish subjects whose behavior is consistent with the deterrence hypothesis and 50 % fair-minded subjects for which intermediate incentives backfire. Since we observe hysteresis, a model of lasting crowding out of fairness concerns explains our data best.

We have contributed to the empirical literature on crowding out in various ways. First, we observe crowding out of fairness concerns in a very simple setting which does not leave a lot of scope for the triggers of crowding out that are usually stressed in this literature. Second, we have established the existence of crowding out as a reaction to probabilistic incentives[24] and in a new domain, namely when incentives are set to deter criminal activities. Third, our comparison of behavior in part 1 with behavior in part 2 provides further evidence for lasting crowding out as it is observed by Irlenbusch and Sliwka (2005), Gneezy and Rustichini (2000), and Gächter, Königstein and Kessler (2005). While it exists for many subjects, we have also observed some subjects whose fairness concerns are - at least partially - reestablished when incentives are reduced or removed completely. The circumstances under which crowding out is lasting remain a topic for future research. Fourth, our study has explicitly focused on the domain of small incentives that are especially important in real life[25]: we have run four out of six treatments with small incentives

[24] To our knowledge the only other paper that documents the existence of crowding out of intrinsic motivation due to probabilistic incentives is Schulze and Frank (2003).

[25] In Germany, the clearance rate for thefts with (without) aggravating circumstances was 14 % (44 %) in 2005 (Polizeiliche Kriminalstatistik, 2005, Table 23). Andreoni et al. (1998) present figures for tax evasion in the US: in 1995, the audit rate for individual tax return was only 1.7 %, the penalty for underpayment of taxes usually 20 % of the underpayment. Polinsky and Shavell (2000b) point out that in general the severity of punishment is quite low in relation what potential

that according to standard neoclassical theory should not deter risk neutral subjects. Thus, we have several treatments to analyze whether crowding out is a continuous or discontinuous process. Our within subject analysis finds evidence for both.

Interestingly, incentives - even in this very simple and plain context - backfire. Kahneman and Tversky's (1986) argument that extrinsic incentives shift the context from an ethical and other-regarding to an instrumental and self-regarding one seems to be adequate for our results. Similarly, the findings confirm those of Houser et al. (2007) who show that crowding out of intrinsic motivation is not only caused by the intentions that incentives signal, but also by incentives *per se*.

What are the policy implications from our experimental study? Taking our data literally would imply to punish criminal behavior either hard or not at all in order to avoid backfiring of small incentives. Of course, the laboratory may abstract from social norms and stigmata that could be "the" driving forces behind punishment in reducing criminal behavior. Thus, we do not conclude that punishment does not work outside the laboratory. However, our data directly reject the deterrence hypothesis that relies on punishment whose effectiveness is caused by pure incentive effects that are independent of all other factors that may influence crime. Our results show that if crime were a gamble - as economists generally argue and as we have modeled it in the laboratory - pure incentives may not work: Especially small and intermediate incentives backfire and may crowd out fairness concerns lastingly. Thus, to convincingly contribute to the discussion on how to efficiently deter crime economists should go beyond the standard deterrence hypothesis.

# 7 Appendix

## 7.1 Experimental sessions and instructions

The order of events during each experimental session was the following: Subjects were welcomed and randomly assigned a cubicle in the laboratory where they took their decisions in complete anonymity from the other participants. The random allocation to a cubicle also determined a subject's role in all three parts. Subjects were handed out the general instructions for the experiment as well as the instructions for part 1. After all subjects had read both instructions carefully and all remaining questions were answered we proceeded to the decision stage of the first part. Part 2 and 3 were conducted in an analogous way. We finished each experimental session by

offenders are capable to pay.

letting subjects answer a questionnaire that asked for demographic characteristics and included a paid Holt and Laury (2002) table. This table was explained in detail in the questionnaire and it was highlighted that one randomly drawn decision from the table was paid out in addition to the earnings in the previous parts.

Instructions, the program, and the questionnaire were originally written in German. The translated general instructions, the translated instructions of the neutrally framed treatment T4 in part 1 for agent $B$, and the translated Holt and Laury (2002) table can be found in the following. Instructions for part 2 and part 3 are as similar to part 1 as possible. For the framed treatments, we used the expression "steal any integer amount between 0 and 90 from participant A" instead of "choose any integer amount between 0 and 90 that shall be transferred from participant A to you", and the term "minus a penalty of $x$ points" instead of "minus an amount of $x$ points".

### 7.1.1 Translated general instructions

General explanations concerning the experiment

Welcome to this experiment. You and the other participants are asked to make decisions. Your decisions as well as the decisions of the other participants will determine the result of the experiment. At the end of the experiment you will be paid **in cash** according to the **actual** result. So please read the instructions thoroughly and think about your decision carefully.

During the experiment you are not allowed to talk to the other participants, to use cell phones or to start any other programs on the computer. The neglect of these rules will lead to the immediate exclusion from the experiment and all payments. If you have any questions, please raise your hand. An experimenter will then come to your seat to answer your questions.

During the experiment we will talk about points instead of Euros. Your total income will therefore be calculated in points first. At the end of the experiment the total amount of points will be converted into Euros according to the following exchange rate:

**1 point = 15 Cents.**

The experiment consists of three **independent** parts in which you can accumulate points. Before each part only the instructions of this part will be handed out.

**During the experiment neither you nor the other participants will receive any information on the course of the experiment (e.g. decisions of other participants or results of a particular part).**

The results of each single part will be calculated only after all three parts will be finished. **Then, one of these three parts will be chosen randomly. At the end of the whole experiment only this part will be paid out in cash according to your decisions.**

### 7.1.2 Translated instructions of the neutrally framed treatment T4 in part 1

Part 1

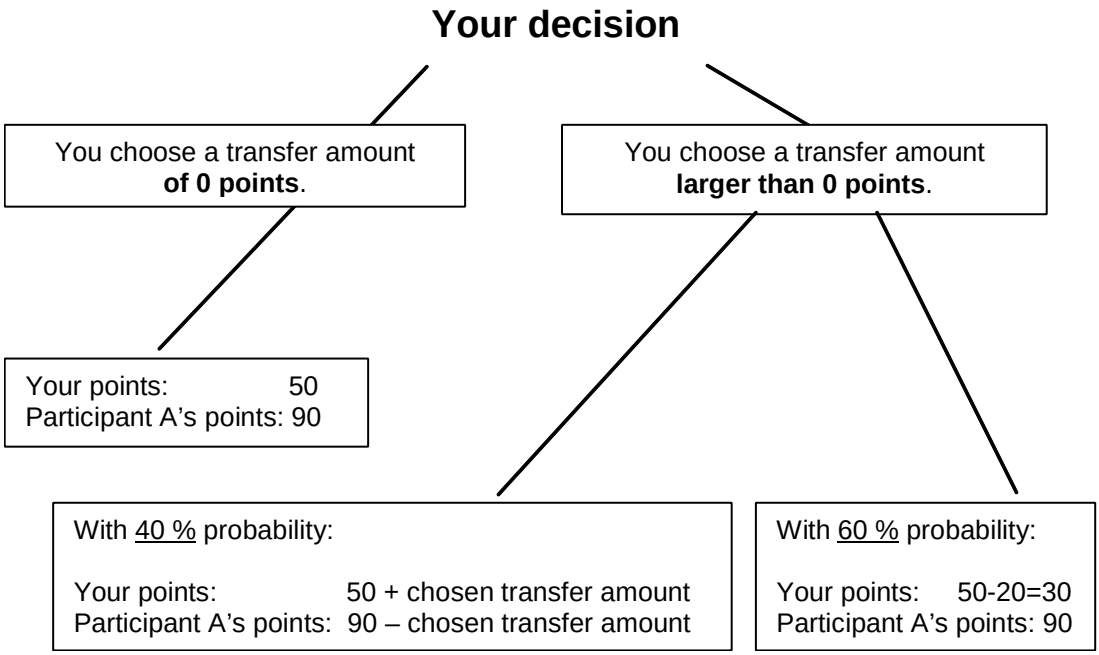
In this part there are **participants in role A and participants in role B. You have been randomly assigned role B for this part. You will be randomly and anonymously matched to another participant in role A.** This random matching lasts only for this part. The matched participant will not be matched to you in the following two parts again. Neither before nor after the experiment will you receive any information about the identity of your matched participant. Likewise, your matched participant will not receive any information about your identity.
As participant B you have an initial endowment of 50 points. Participant A has an initial endowment of 90 points.
As a participant in role B you can choose any **integer amount** between 0 and 90 points (including 0 and 90) **which shall be transferred from participant A to you**. Participant A does not make any decision. In order to make your decision please enter your chosen amount on the corresponding computer screen and push the OK button.

- If you choose a transfer amount of 0 points, you will receive your initial endowment of 50 points, and participant A will receive his initial endowment of 90 points.
- If you choose a transfer amount larger than 0 points,
  - **with *40 %* probability you will receive your initial endowment of 50 points plus your chosen transfer amount and participant A will receive his initial endowment of 90 points minus your chosen transfer amount.**
  - **with *60 %* probability you will receive your initial endowment of 50 points minus an amount of 20 points, i.e. 30 points, and participant A will receive his initial endowment of 90 points.**

*Example 1*: You choose a transfer amount of 22 points. With *40 %* probability you will receive 50 + 22 points = 72 points, and with *60 %* probability you will receive 50 – 20 points = 30 points. Participant A will receive 90 – 22 points = 68 points with *40 %* probability and his initial endowment of 90 points with *60 %* probability.

*Example 2*: You choose a transfer amount of 0 points. You will receive 50 points. Participant A will receive 90 points.

The course of action of part 1 is illustrated by the following graph:




If you have any questions, please raise your hand. An experimenter will come to your seat to answer your questions.

### 7.1.3 Translated Holt and Laury (2002) table

| Decision | Option A | Option B |
|---|---|---|
| Decision 1 | 10 points | 25 points with a probability of 10 %<br>0 points with a probability of 90 % |
| Decision 2 | 10 points | 25 points with a probability of 20 %<br>0 points with a probability of 80 % |
| Decision 3 | 10 points | 25 points with a probability of 30 %<br>0 points with a probability of 70 % |
| Decision 4 | 10 points | 25 points with a probability of 40 %<br>0 points with a probability of 60 % |
| Decision 5 | 10 points | 25 points with a probability of 50 %<br>0 points with a probability of 50 % |
| Decision 6 | 10 points | 25 points with a probability of 60 %<br>0 points with a probability of 40 % |
| Decision 7 | 10 points | 25 points with a probability of 70 %<br>0 points with a probability of 30 % |
| Decision 8 | 10 points | 25 points with a probability of 80 %<br>0 points with a probability of 20 % |
| Decision 9 | 10 points | 25 points with a probability of 90 %<br>0 points with a probability of 10 % |
| Decision 10 | 10 points | 25 points with a probability of 100 %<br>0 points with a probability of 0 % |

Participants made 10 separate decisions whether they preferred option A to option B. Option B varied by the decisions with the associated probabilities displayed above. One decision was chosen randomly (all with equal probability) and paid at the end of the experiment.

We classify the observed 51 subjects who prefer option A to option B in decisions 1 to 4 and option B to option A otherwise as risk-neutral. The observed 16 subjects preferring option A in decisions 1 to 5 are categorized as risk-seeking. We observe 88 risk-averse subjects indicating option A in decisions 1 to $k$, with $k > 5$. Three subjects behave irrationally (or are humble) in the sense that they prefer option A to option B in decision 10.

# References


[1] Andreoni, J., Erard, B., Feinstein, J., 1998. Tax compliance. Journal of Economic Literature 36, 818-860.

[2] Becker, G. S., 1968. Crime and punishment, an economic approach. Journal of Political Economy 76 (2), 169-217.

[3] Bohnet, I., Cooter, R. D., 2005. Expressive law, framing or equilibrium selection? Mimeo.

[4] Bowles, S., 2007. Social preferences and public policies, are good laws a substitute for good citizens?. University of Siena working paper no. 496.

[5] Bolton, G. E., Ockenfels, A., 2000. A theory of equity, reciprocity and competition. American Economic Review 90 (1), 166-193.

[6] Bundeskriminalamt (Ed.), 2005. Polizeiliche Kriminalstatistik Bundesrepublik Deutschland, Berichtsjahr 2005, Wiesbaden, http://www.bka.de/pks/pks2005/download/pks-jb_2005_bka.pdf.

[7] Deci, E. L., 1971. Effects of externally mediated rewards on intrinsic motivation. Journal of Personality and Social Psychology 18 (1), 105-115.

[8] Deci, E. L., Koestner, R., Ryan, M. R., 1999. A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. Psychological Bulletin 125 (6), 627-668.

[9] Eide, E., 2000. Economics of criminal behavior. In: Bouckaert, B., De Geest, G. (Eds.), Encyclopedia of Law and Economics, Vol. V, Cheltenham, Edward Elgar, 345-89.

[10] Falk, A., Fischbacher, U., 2002. Crime in the lab, detecting social interaction. European Economic Review 46 (4-5), 859-869.

[11] Fehr, E., Falk, A., 2002. Psychological foundations of incentives. European Economic Review 46 (4-5), 687-724

[12] Fehr, E., Schmidt, K. M., 1999. A theory of fairness, competition, and cooperation. Quarterly Journal of Economics 114 (3), 817-868.

[13] Fehr, E., Schmidt, K. M., 2006. The economics of fairness, reciprocity and altruism - Experimental evidence and new theories. In: Kolm, S.-C., Ythier, J. M. (Eds.), Handbook of the Economics of Giving, Altruism and Reciprocity, Vol. 1, Amsterdam, Elsevier, 615-91

[14] Forsythe, R., Horowitz, J. L., Savin, N. E., Sefton, M., 1994. Fairness in simple bargaining experiments. Games and Economic Behavior 6 (3), 347-369.

[15] Frey, B. S., Jegen, R., 2001. Motivation crowding theory. Journal of Economic Surveys 15 (5), 589-611.

[16] Frey, B. S., Oberholzer-Gee, F., 1997. The costs of price incentives, an empirical analysis of motivation crowding-out. American Economic Review 87 (4), 746-755.

[17] Gächter, S., Kessler, E., Königstein, M., 2006. Performance incentives and the dynamics of voluntary cooperation, an experimental investigation. Mimeo.

[18] Galbiati, R., Vertova, P., 2005. Law and behaviours in social dilemmas, testing the effect of obligations on cooperation. Working paper.

[19] Garoupa, N., 1997. The theory of optimal law enforcement. Journal of Economic Surveys 11 (3), 267-295.

[20] Glaeser, E. L., 1999. An overview of crime and punishment. Mimeo.

[21] Gneezy, U., 2003. The w effect of incentives. Mimeo.

[22] Gneezy, U., Rustichini, A., 2000a. A fine is a price. Journal of Legal Studies 29 (1), 1-17.

[23] Gneezy, U., Rustichini, A., 2000b. Pay enough or don't pay at all. Quarterly Journal of Economics 115 (3), 791-810.

[24] Holt, C. A., Laury, S. K., 2002. Risk aversion and incentive effects. American Economic Review 92 (5), 1644-1655.

[25] Houser, D., Xiao, E., McCabe, K., Smith, V., 2007. When punishment fails, research on sanctions, intentions and non-cooperation, Games and Economic Behavior, forthcoming.

[26] Irlenbusch, B., Sliwka, D., 2005. Incentives, decision frames, and motivation crowding out, an experimental investigation. IZA DP no. 1758.

[27] Kahneman, D., Tversky, A., 1986. Rational choice and the framing of decisions. Journal of Business 59 (4), 251-278.

[28] Lepper, M. R., Greene, D., Nisbett, R. E., 1973. Undermining children's intrinsic interest with extrinsic reward, a test of the "overjustification" hypothesis. Journal of Personality and Social Psychology 28 (1), 129-137.

[29] Levitt, S., 1997. Using electoral cycles in police hiring to estimate the effect of police on crime. American Economic Review 87(3), 270-290.

[30] Polinsky, A. M., Shavell, S., 2000a. The economic theory of public enforcement of law. Journal of Economic Literature 38 (1), 45-76.

[31] Polinsky, A. M., Shavell, S., 2000b. The fairness of sanctions, some implications for optimal enforcement theory. American Law and Economics Review 2 (2), 223-237.

[32] Schulze, G. G., Frank, B., 2003. Deterrence versus intrinsic motivation, experimental evidence on the determinants of corruptibility. Economics of Governance, 4, 143-160.

[33] Torgler, B., 2002. Speaking to theorists and searching for facts, tax morale and tax compliance in experiments. Journal of Economic Surveys 16 (5), 657-683.

[34] Trautmann, S. T., 2007. Fehr-Schmidt process fairness and dynamic consistency. Working paper.

[35] Tyran, J.-R., Feld, L. P., 2006. Achieving compliance when legal sanctions are non-deterrent. Scandinavian Journal of Economics 108 (1), 135-156.