Turning the Lab into Jeremy Bentham’s Panopticon
The Effect of Punishment on Offenders and Non-Offenders

Christoph Engel
Bernd Irlenbusch
Turning the Lab into Jeremy Bentham’s Panopticon
The Effect of Punishment on Offenders and Non-Offenders

Christoph Engel / Bernd Irlenbusch

February 2010
Abstract

The most famous element in Bentham’s theory of punishment, the Panopticon Prison, expresses his view of the two purposes of punishment, deterrence and special prevention. We investigate Bentham’s intuition in a public goods lab experiment by manipulating how much information on punishment experienced by others is available to would-be offenders. Compared with the tone that Jeremy Bentham set, our results are non-expected: If would-be offenders learn about contributions and punishment of others at the individual level, they contribute much less to the public project. Our results confirm the special prevention effect but show that the deterrence effect is smaller the more information on individual punishment is available.

Keywords: Punishment, Deterrence, Special Prevention, Jeremy Bentham, Experiment, Public Good

JEL: C91, H41, K14, K42

Research assistance by Karsten Lorenz and Lilia Zhurakhovska and helpful comments by Markus Englerth and Andreas Nicklisch are gratefully acknowledged.
I. Research Question

Not many scientific achievements are cast in stone. Jeremy Bentham’s theory of punishment is among the rare exceptions. Inspired by his younger brother Samuel, a naval architect based in Russia in the 1780s, he translated his theory into a blueprint for the design of prisons (Semple 1993). Over time, all over the world some 20 "panopticon" prisons have been built. The basic idea is simple. The circular construction places each and every inmate under permanent control by the supervisor, located at the centre. Every spectator cannot but realize that prisoners have fully lost their autonomy.

![Jeremy Bentham's Design of the Panopticon](image)

The architecture implements the purpose punishment is supposed to serve (Bentham 1830:Book V Chapter III):

“Hence the prevention of offenses divides itself into two branches: Particular prevention, which applies to the delinquent himself; and general prevention, which is applicable to all the members of the community without exception.

General prevention is effected by the denunciation of punishment, and by its application, which, according to the common expression, serves for an example. The punishment suffered by the offender presents to everyone an example of what he himself will have to suffer if he is guilty of the same offense.

General prevention ought to be the chief end of punishment, as it is its real justification. [...] The punishment inflicted on the individual becomes a source of security to all. [...] That punishment, which, considered in itself, appeared base and repugnant to all generous sentiments, is elevated to the first rank of benefits, when it is regarded

---

not as an act of wrath or of vengeance against a guilty or unfortunate individual who has given way to mischievous inclinations, but as an indispensable sacrifice to the common safety” (Bentham 1830:Book I Chapter III).

In this paper we investigate Bentham’s intuition of special prevention and general prevention (deterrence) in a lab experiment. A lab experiment has the decisive advantage to give us full control over the institutional setting. We can therefore manipulate how much information on punishment experienced by others is available. Additionally, we can unambiguously see whether and to which degree subjects are well-behaved, and whether or not they change their behavior after having observed or experienced punishment. Our “work-horse” is a standard public goods game, that has already been extensively investigated (Ostrom, Walker et al. 1992; Fehr and Gächter 2000; Fischbacher and Gächter 2009). Four players interact over 10 periods. In each period they can contribute to a public good. The individual return of each player from one unit of contribution, i.e. the marginal capita return, is such that from an individual perspective it is unprofitable to contribute. However, since all players benefit, full contribution would maximize total profit. In a second stage of each round, a fifth player observes the individual contributions and, based on this information, can punish each of the four players who can make contributions. Punishing is costly for the fifth player, i.e., she must invest her own money if she decides to diminish the payoff of one or more of the others. All five players individually have the same return from the public good irrespective of whether or not and how much they contributed.

To investigate Jeremy Bentham’s idea, we manipulate feedback. In the low treatment, contributors only learn aggregate contributions. In the medium treatment, they also learn aggregate punishment. In the high treatment, they know individual contributions and individual punishment. As a further test, we invite another group of four contributors for another 10 periods. The group supervisor stays in office. Before the second group starts playing, graphs inform them about their predecessor’s performance. The information about the contributions and, if applicable, about punishment in the first 10 periods is the same as was given to contributors during the first phase. Also the degree of feedback is kept constant across phases. For instance if feedback was low in phase 1, the group of successors is not informed about punishment either, neither with respect to their predecessors, nor with respect to other players of the current group.

Compared with the tone that Jeremy Bentham set, we have a non-expected result: punishment information is at best immaterial. If bystanders learn both behavior and correctional responses at the individual level, they contribute much less to the public project. However, punishment does not miss its intended effect, neither on those punished themselves, nor on bystanders. Yet the effect is indirect. Bystanders and newly arrived group members are mainly influenced by the observed previous contribution levels in their groups that, of course, have been affected by punishment.
II. Related Literature

Behavior in our work-horse setting exhibits a robust and well-known pattern. In the absence of punishment, players contribute significantly in the beginning, but contributions decay quickly (for summaries see Ledyard 1995; Zelmer 2003). More recent studies show that contributions stabilize if subjects are given a chance to punish each other after having observed individual contributions (for a summary see Herrmann, Thöni et al. 2008). In this literature, punishment is typically decentralized. Few experimental studies use centralized punishment, but they focus on different research questions. For example, they investigate whether unaffected outsiders are willing to spend money for disciplining others (Fehr and Fischbacher 2004), how leaders can motivate members of their team by the threat of punishment (Güth, Levati et al. 2007; Güürer, Irlenbusch et al. 2009), whether centralized punishment develops endogenously when players can voluntarily join a sanctioning institution (Kosfeld, Okada et al. 2008), how centralized punishment affects behavior in a threshold public goods game (Guillén, Schwieren et al. 2006), and how mild central sanctions interact with social norms (Tyran and Feld 2006; Galbiati and Vertova 2008b; Galbiati and Vertova 2008a). Putterman, Tyran, and Kamei (2010) are able to choose efficient formal sanctions by voting.

Very few experimental studies investigate the impact of information about others’ behavior (also called “social history”) on own behavior. (Berg, Dickhaut et al. 1995) find that providing a social history increases cooperation in their trust game setting. (Fehr and Rockenbach 2003), however, do not find a change in subjects’ behavior in a gift exchange game with punishment. Informing responders about the average offers before they decide whether to accept or reject their specific offer seems to significantly increase offers and offer-specific rejection probabilities (Bohnet and Zeckhauser 2004). In a binary dictator game (Krupka and Weber 2009) find that showing subjects what others actually do produces more pro-social behavior. Interestingly, this is even the case when observed subjects are mostly selfish. They also find support for an informational effect: observing more people behaving pro-socially generally produces more pro-social behavior. We are not aware of any study that investigates the influence of information about others’ behavior in a public good setting with punishment.

There is a growing body of experiments in criminology (for summaries see Farrington 2003; Nagin and Pogarsky 2003; Petrosino, Turpin-Petrosino et al. 2003; Farrington and Welsh 2005; Farrington and Welsh 2006; Petrosino, Kiff et al. 2006). Many are quasi experiments in the field (Farrington and Welsh 2006). Apparently though no experiment has tried to assess the effect of punishment on true outsiders to the criminal system (cf. the comprehensive survey by Farrington and Welsh 2006).

In modern parlance, Jeremy Bentham proposes deterrence. The deterrence literature is rich (for summaries see Cameron 1988; Ehrlich 1996; Nagin 1999b; Nagin and Pogarsky 2001; Sherman 2002; Levitt 2004; Matsueda, Kreager et al. 2006). There is widespread consensus that punishment has an effect on non-criminals (Andenaes 1974; Gibbs 1979; Katyal 1997:2444; Frase 2005:69). Specifically the social benefit from punishment is not confined to incapacitation, i.e. to

Much empirical effort has been devoted to showing that and how the objective certainty, celerity and severity of sanctions is modulated by perception (Williams and Hawkins 1986; Paternoster 1987; Nagin 1998b; Anderson 2002; Piquero and Pogarsky 2002; Pogarsky, Piquero et al. 2004; Wright, Caspi et al. 2004; Matsueda, Kreager et al. 2006). The difference can be pronounced. For instance, one interview study with prison inmates showed that 76% of active criminals, and 89% of most violent criminals, either perceive no risk of apprehension or are incognizant of the likely punishment (Anderson 2002:295).

One particularly troubling finding has been called the resetting effect. Empirically, those who have been punished in the past not so rarely increase rather than decrease criminal behavior (Piquero and Paternoster 1998; Piquero and Pogarsky 2002; Pogarsky and Piquero 2003). A prominent explanation for this observation is Bayesian updating. Through the one experience with the judicial system, offenders learn that the probability of arrest, conviction and enforcement is smaller than expected, and that they suffer less from the sanction than they thought (Paternoster, Saltzmann et al. 1983; Nagin 1998b:15 f.). If this experience spreads, the collateral damage of criminal justice is even larger. Also criminals cannot be stigmatized twice, which is why the opportunity cost of the second sanction is lower (Harel and Klement 2007). Through stigma, the prospect for earnings from legal occupation have decreased (Schmidt and Witte 1984).

### III. Hypotheses

Our first hypothesis translates Jeremy Bentham’s proposal into our setting:

H$_1$: The more punishment is visible, the higher are contributions.

From the more recent literature in criminology, we derive

H$_2$: The more punishment is severe, and the more it is consistent, the higher are contributions.

Jeremy Bentham’s proposal is rooted in his conviction that would-be criminals are also sensitive to foreign experiences with the criminal system. This triggers
H3: Those not punished in the past contribute at least as much to the public project as those punished. More detailed information about contributions and punishment of predecessors has a favorable effect.

IV. Experimental Design

324 students (149 female) from a variety of majors participated in an experiment conducted at the Econ Lab of Cologne University in December 2008. The experiment was implemented in zTree (Fischbacher 2007). Participants were invited using Orsee (Greiner 2003) and were randomly assigned to treatments.2

Participants played ex ante announced 10 rounds of a standard public goods game in anonymous groups of four (called “players of type A” in the instructions) that stayed together during the entire experiment. Per period, each participant received an endowment of 20 talers. Players of type A could decide how many talers to invest in a project. Each taler contributed to the project created a marginal per capita return of 0.4. Before the start of the game, per group one additional subject was randomly assigned as the group supervisor whom we called “player of type B” in the instructions. In each round supervisors were informed about the individual contributions of each type A-player. Supervisors had the same endowment and the same marginal per capita return from the project. However, they could not contribute. Rather they could spend their endowment on individually punishing the type A-players. For type B players, contributions of type A players thus are the equivalent of a levy from which a public official is financed. Following Fehr and Gächter (2002), the punishment technology was linear: one taler invested for punishment destroyed three talers of the punished player.

Our experiment consisted of two phases. The first group of four type A players in phase 1 was followed by a second phase with a fresh group of four subjects. Also the second phase consisted of ex ante announced 10 rounds. Only the supervisor stayed the same in both phases. Before starting to play themselves, the second group received graphs informing them about the performance and/or received punishment in their respective group of predecessors.

Our three treatments differed in the degree of feedback for type A-players. An overview of the differences in feedback is provided in Table 1. In each phase feedback is provided about previous rounds, and in phase 2 also about play of the respective group in phase 1. In treatments low and medium, the supervisor cannot identify players of type A across periods. By contrast, in treatment high, players of type A are identified by numbers that stay constant over all rounds.

---

2 The translation of the instructions for one of our treatments are provided in the appendix. Original instructions were in German. All instructions can be obtained on request from the authors.
<table>
<thead>
<tr>
<th>Treatment</th>
<th>feedback about previous rounds</th>
<th>feedback about all rounds in previous phase (in phase 2 only)</th>
</tr>
</thead>
<tbody>
<tr>
<td>low</td>
<td>• average contributions</td>
<td>• average contributions</td>
</tr>
<tr>
<td>medium</td>
<td>(in addition to information provided in low)</td>
<td>• average received punishment</td>
</tr>
<tr>
<td>high</td>
<td>(in addition to information provided in medium)</td>
<td>• individual contributions • individual earnings • individual received punishment</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• average contributions</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• average received punishment</td>
</tr>
</tbody>
</table>

Table 1
Feedback Provided to Type A-Players in Different Treatments

Of course, in a number of respects, our design differs from the reality of criminal policy. Stakes are much lower, both for society and for the “offender”. But we keep the basic dilemma structure that also underlies most criminal offenses: the criminal is best off if she ignores the harm she inflicts on other members of society. Also there is no criminal code. Norms are implicit in our setting. They result from the sanctioning policy of the supervisor. Yet in criminal policy, a related effect is not uncommon: the criminal authorities use the degrees of freedom they dispose of to flexibly react to crime, despite the fact that, at face value, criminal offenses are precisely defined in the respective penal code. Finally, in our experiment, sanctioning is not accompanied by words that express social disapproval or moral indignation. In the experiment, all value judgement is through correctional action. The effect of the negative incentive is less visibly backed up by appealing to the offender’s identity as a member of this one society.

Yet in recompense for the slight reduction of external validity, our design gives us control over elements that have plagued criminal policy and empirical research alike. Our supervisor benefits directly from the contributions of group members to the project. They cannot personally identify group members so that repeated game effects after the experiment are not an issue. Effects of retribution can be excluded (cf. Wood 2002). Incapacitation is impossible in our setting (cf. Kessler and Levitt 1999). Intervention cannot shift criminal activity to another location (cf. Hakim, Spiegel et al. 1984). Victims cannot respond by moving to a different town or quarter (cf. Anderson 1990). Players are perfectly symmetric, so that bystanders have no reason to expect that they will be treated any differently if they behave the same way (cf. Robinson and Darley 2003:973). Since each round only lasts minutes, arguably discounting should be negligible (cf. Levitt 1998:353).

Given our completely neutral framing and decontextualisation and the fact that there cannot be competing tasks, we need not be too much concerned about impulsivity (cf. Shepherd 2004) or about crime as symbol (cf. Matsueda, Kreager et al. 2006:103). The risk of sanction misperception (Nagin 1998b:19) is minimized. Subjects are perfectly informed about punishment inflicted on themselves and, depending on our treatments, on others. Arguably, habituation does not mat-
term, given that players only play 10 rounds, and that the entire experiment lasts little longer than an hour (cf. Hawkins 1969:560). Again due to anonymity, the fact that would-be offenders are members of a peer group with criminal propensity cannot explain behavior (cf. Kahan 1997a:2486). Punishment cannot serve as a "badge of honor" (cf. Wilson and Herrnstein 1985:304). Moral credibility (cf. Kahan 1997a:2481) and the mirror concept of moral condemnation (cf. Kahan 1997b:383) cannot matter either. We deliberately do not speak of "punishment", but only of a "reduction of income". The supervisor is neutrally labeled as a "player of type B". Thereby we reduce the expressive function of punishment (cf. Kahan 1997a:2483) to a minimum. Since we guarantee anonymity, formal sanctions cannot be supplemented or complemented by informal sanctions in treatments low and medium (cf. Cameron 1988:302). In treatment high, free-riders are labeled (cf. Lemert 1951; Becker 1963). But other group members do not have an opportunity for a targeted reaction.

Our treatments discriminate between several explanations for the effect of punishment. The treatments are directly targeted to sanction risk perception (cf. Nagin 1998b:1, criticizing the absence of empirical evidence on this). In treatment low, subjects do not get feedback about sanctions directed to other group members. Consequently participants not punished themselves in the respective period (whom, as a shorthand, we call loyals) can at best be indirectly influenced by the sanctioning policy of their respective supervisor, through its effect on contributions. This is different in treatment medium. However, loyals still do not learn the precise response of the supervisor to a specific deviation from the (implicit) group norm. This information is only available in treatment high.

We can precisely disentangle personal and vicarious experiences (cf. Stafford and Warr 1993). In treatment low, vicarious experiences are excluded by design. In treatments medium and high, in the first round of phase 2, all experiences are vicarious. In all other rounds of both phases, culprits combine personal with vicarious experiences, whereas loyals make differently precise vicarious experiences only. This makes it also possible to test for the resetting effect (cf. Pogarsky and Piquero 2003).

In two complementary ways, we can analyze the effect of ambiguity and uncertainty (cf. Nagin 1998b:11). In treatment low, loyals are completely uncertain about the sanctioning policy of their supervisor. In treatments medium and high, in phase 2 players can adjust their behavior to the sanctioning policy from phase 1. Also in treatments medium and high, we can test how the consistency of the respective supervisor (cf. Katyal 1997:2451) plays itself out in contribution rates. If there is stigmatization (Harel and Klement 2007), this should become visible in comparing treatments medium and high. For only in treatment high, culprits are individualized with a reputation over rounds.

A final advantage of our design results from the character of our dependent variable. In the field, and in criminological research on recidivism for that matter, it usually is binary: either a person abides by the law, or she breaks it. In our experiment, contributions are (quasi) continuous, on a scale from 0 to 20. We therefore cannot only investigate whether those who have been punished,
or those who observe how others are punished, respect the law in the future. We can also measure how much they contribute to the public project. Along with this, we get a meaningful dynamic measure. We can investigate, by how much a participant changes her contribution level after punishment, or upon observing punishment in action.

Participants received a show up fee of 2.50 €. Theoretically, subjects could make real losses. Since the lab has built a reputation that subjects do not put their own money at risk, we gave them an extra 50 talers at the beginning of the experiment, explicitly motivated to cater for potential losses. Earnings were individually and anonymously paid out to all participants at an exchange rate of 0.04 € per taler. On average, total earnings of contributors were 22.23 € (sd 2.64, range [12.07, 33.3]). Total earnings of supervisors, who played 20 periods each, were on average 45.60 € (standard deviation 5.51, range [34.36, 56.71]).

V. Results

1. Anticipation

In the first period of the first phase, active players have no own or vicarious experiences with the respective punishment institution. But they are fully informed about institutional design. Is this information sufficient to induce different behavior in different treatments? Is a potential effect of the institution anticipated? Although descriptive figures point into the direction of the treatment effects, if we only consider data from the first period we do not find a significant effect of the treatments, neither if we test non-parametrically nor if we test parametrically.

2. Phase 1

a. Full Transparency Reduces Contributions

Contributions and deductions through punishment were as in Figure 2 and in Table 2. In the Appendix, we also provide graphs for individual groups. Our main result is visible immediately: full transparency hurts, while partial transparency is immaterial. In treatment high, absolute contributions are lower than in the remaining two treatments. The difference in absolute contributions results from less favorable contribution dynamics. While contributions rise quickly in treatments low and medium, they remain almost stable in treatment high. Visual inspection suggests that

---

3 Mean contributions are low 14.521, medium 14.563, high 13.208.
4 Mann Whitney, low vs. medium, N = 96, p = .932; low vs. high, p = .2955; medium vs. high, p = .2514. All tests are two-sided. Note that in the first period individual contribution decisions are still fully independent of each other.
5 In the first period, 61 out of 144 (active) participants contribute their entire endowment of 20 taler. 4 keep the whole endowment for themselves. This data structure makes a Tobit model appropriate. In this model, treatment high is the reference category. Regressors for treatments low (p = .229) and medium (p = .184) are independently and jointly insignificant (Wald test, F(2, 142) = 1.09, p = .3376).
informing participants about average punishment, i.e. the difference between treatments low and medium, is close to irrelevant.

<table>
<thead>
<tr>
<th>Phase</th>
<th>first</th>
<th>second</th>
</tr>
</thead>
<tbody>
<tr>
<td>treatment</td>
<td>contr</td>
<td>recpun</td>
</tr>
<tr>
<td>Low</td>
<td>17.45</td>
<td>1.7</td>
</tr>
<tr>
<td>Medium</td>
<td>16.99</td>
<td>2.68</td>
</tr>
<tr>
<td>high</td>
<td>14.02</td>
<td>2.84</td>
</tr>
</tbody>
</table>

Table 2
Means of Contributions and Received Punishment

For this and all later non-parametric tests, we consider means over the first ten rounds. We report p-values from two-sided tests. If we compare treatments, we use a Mann-Whitney test. In this test, the difference between low and high is indeed weakly significant (N = 24, p = .0647), while the remaining comparisons are insignificant. In the Appendix, we explain our strategy for parametric estimation. Using this strategy, in the regression of Table 3 we establish significantly higher contributions in the low and medium treatments, compared to treatment high, which serves as the reference category; the endgame effect is captured by the regressor for the final period. Full feedback (high) thus reduces contributions. We squarely refute H1: contributions are lower, the more punishment information is available.

Figure 2
Summary Results

---

6 On the mirror model that ignores censoring and clustering, the Hausman test is insignificant, which indicates that we are justified to prefer the more efficient random effects over the fixed effects model.
<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>low</td>
<td>8.185**</td>
</tr>
<tr>
<td>medium</td>
<td>6.749*</td>
</tr>
<tr>
<td>final period</td>
<td>-1.283</td>
</tr>
<tr>
<td>cons</td>
<td>16.847***</td>
</tr>
<tr>
<td>N</td>
<td>1440</td>
</tr>
<tr>
<td>left censored</td>
<td>52</td>
</tr>
<tr>
<td>right censored</td>
<td>784</td>
</tr>
<tr>
<td>p model</td>
<td>&lt;.001</td>
</tr>
</tbody>
</table>

Table 3
Parametric Test of Treatment Effects
depvar: contribution
random effects Tobit, bootstrapped at the group level, 500 reps
*** p < .001, ** p < .01, * p < .05

Making average punishment explicit (medium) does not significantly improve contributions, as shown by a Wald test of the null hypothesis that coefficients for low and medium are the same (p = .6096).

b. Impression Management is Key

As the models of Table 4 show, contributions in the next periods are chiefly explained by contributions in the respective previous periods. In both models, apart from the regressor for the endgame effect, this is the only regressor that is significant at conventional levels. Actually it is highly significant. Those who have not been punished in the previous period contribute more, but this effect is only weakly significant, model 2. This is not just a sign of stickiness or conservatism. As has been shown, in motivational terms groups tend to be heterogeneous. There is a small fraction of true altruists, and a somewhat larger fraction of profit maximizing egoists. Yet most group members are conditional cooperators. They are happy to contribute (much) to the public project, as long as they perceive other group members to contribute (much) as well (Fischbacher, Gächter et al. 2001; Fischbacher and Gächter 2009). Conditional cooperators need not be deterred. But they care about the risk of being “the sucker”. Hence they care about a signal of expected cooperativeness in their group. As the regressions of Table 4 demonstrate, average contributions in the previous round serve as this signal.

---

7 For both models, if one conducts the Hausman test on the mirror model that ignores censoring and clustering, it is significant. The problem can, however, be remedied by a Hausman Taylor model, see Appendix for methodology. In this model, we treat regressor “final” as exogenous.
<table>
<thead>
<tr>
<th>not punished in previous period</th>
<th>model 1</th>
<th>model 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average contributions in previous round</td>
<td>1.502***</td>
<td>1.475***</td>
</tr>
<tr>
<td>Low</td>
<td>5.224</td>
<td>5.365</td>
</tr>
<tr>
<td>Lavcontr*low</td>
<td>-.140</td>
<td>-.153</td>
</tr>
<tr>
<td>Medium</td>
<td>1.540</td>
<td>1.798</td>
</tr>
<tr>
<td>Lavcontr*medium</td>
<td>.041</td>
<td>.029</td>
</tr>
<tr>
<td>final period</td>
<td>-3.155***</td>
<td>-3.241***</td>
</tr>
<tr>
<td>Cons</td>
<td>-4.272*</td>
<td>-4.515*</td>
</tr>
<tr>
<td>N</td>
<td>1296</td>
<td>1296</td>
</tr>
<tr>
<td>p model</td>
<td>&lt;.001</td>
<td>&lt;.001</td>
</tr>
</tbody>
</table>

### Table 4
Effect of Observed Cooperativeness
depvar: contribution
Lavcontr: average contribution in previous period
random effects Tobit, bootstrapped at the group level, 500 reps
*** p < .001, ** p < .01, * p < .05, + p < .1

**c. Starting a Virtuous Cycle**

Supervisors in our experiment seem to try to establish a norm of high cooperation right from the beginning. Inspecting Figure 2, one realizes that contributions in treatments *low* and *medium* on the one hand and treatment *high* on the other hand differ most in the beginning. In all treatments, there is quite some punishment in early rounds. The amount of punishment in period 1 does not significantly differ between treatments, neither in a nonparametric (Mann Whitney, N = 24, *low* vs. *high*, p = .8165; *low* vs. *medium*, p = .6215; *medium* vs. *high*, p = .8391) nor in a parametric test.\(^8\) However, the substantial intervention works well in *low* and *medium*, while it is only mildly effective in *high*. Actually, in *medium*, active players do not immediately adjust to the implicit group norm. It takes one more period until contributions jump up. Yet if one compares the difference between period 1 and period 3 contributions, at the group level one establishes a significant difference between *low* and *high* (Mann Whitney, N = 24, p = .0119) and a weakly significant difference between *medium* vs. *high* (p = .0937).

**d. The Indirect Effect of Punishment**

Why does punishment seem to work better if individual contributions and punishment are less visible? Let us first show that punishment works. If it works, it induces punishees to increase their contributions. We therefore work with first differences of contributions as the dependent variable. In the regression of Table 5, the dummy for not having been punished in the previous

---

\(^8\) If one regresses average punishment per group on treatment dummies, the dummies are insignificant, *low* p = .723, *medium* p = .737.
period is highly significant ($z = -5.69$) and strongly negative.\(^9\) This implies that those who have been punished are much more likely to increase their contributions than those who have not.

However, both in low and in medium, the treatment effect and its interaction with not having been punished in the previous period more or less cancel out. Since the negative main effect of not having been punished is bigger than the positive constant, in all three treatments the model predicts that those not punished slightly reduce their contributions (-.439 in low, -.625 in medium, -.530 in high). Wald tests show that, for those not punished in the previous period, there is no difference between treatments (low vs. medium, $p = .5790$; low vs. high, $p = .8031$; medium vs. high, $p = .8110$).\(^10\) Consequently, in a dynamic perspective, the difference between treatments does not originate in the reaction of those who have not been punished.

The critical difference is the reaction of those who have received punishment. In all treatments they increase their contributions. However, the increase is the bigger the less punishment is transparent (4.184 in low, 2.982 in medium, 1.740 in high); the treatment effect adds to the constant. The more punishment is transparent, the less it works on those to whom it has been addressed.

<table>
<thead>
<tr>
<th>not punished in previous period</th>
<th>-2.271***</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>2.244**</td>
</tr>
<tr>
<td>low*Linnoc</td>
<td>-2.353*</td>
</tr>
<tr>
<td>Medium</td>
<td>1.242*</td>
</tr>
<tr>
<td>medium*Linnoc</td>
<td>-1.337*</td>
</tr>
<tr>
<td>final period</td>
<td>-1.217**</td>
</tr>
<tr>
<td>Cons</td>
<td>1.740***</td>
</tr>
<tr>
<td>N</td>
<td>1296</td>
</tr>
<tr>
<td>$p$ model</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>$R^2$ within</td>
<td>.1959</td>
</tr>
<tr>
<td>$R^2$ between</td>
<td>.1164</td>
</tr>
<tr>
<td>$R^2$ overall</td>
<td>.1562</td>
</tr>
</tbody>
</table>

Table 5
Explaining Changes in Contributions
Random effects, robust standard errors, clustered for groups
Linnoc: not punished in previous period

\(*** p < .001, ** p < .01, * p < .05\)

To appreciate this finding, it is informative to separately consider those who have and who have not been punished in the previous period, which we do in Figure 3 for the critical first three periods. In all treatments, those who have been punished are more likely to increase their contribu-

---

\(^9\) Data on first differences is not censored, so that there is no need for Tobit. The Hausman test on the mirror model that ignores clustering turns out significant. But the problem can be remedied by a Hausman Taylor model, again treating regressor “final” as exogenous.

\(^10\) To illustrate methodology: for the first comparison, we test the null hypothesis (linnoc +) medium + medium*linnoc = (linnoc +) low + low*linnoc.
tions than those who have not (Mann Whitney, N = 68, p = .0029). Most of those not punished do not change their contribution level at all. If they had made substantial contributions in the first round, this is good news. If, however, first-round contributions were low, it is bad policy for the supervisor not to intervene. Most importantly, it is very rare in low for a punished participant to reduce her contributions in response. This is more frequent in medium, and much more frequent in high. In criminological parlance: the higher punishment transparency, the more pronounced the resetting effect.

![Figure 3: Initial Dynamics](image)

Figure 3
Initial Dynamics


indepvar: dummy that is 1 if the participant has been punished in period 1 (for the first change) and in period 2 (for the second change)

e. **Why Does Transparency Make Punishment Less Effective?**

We therefore must look out for reasons why higher transparency might make punishment less effective. A first hint is provided by the regression in Table 6. The more a participant has been punished in the previous period the less she contributes in the following period. Since we have

---

11 Two datapoints per group, one is the average of those who have not been punished, the other is the average of those who have been punished. Punishment is considered for periods 1 and 2, the dependent variable for periods 2 and 3. N < 72 results from the fact that in 3 groups nobody was punished, and in 1 group all were punished.

12 The difference between low and high is significant, Mann Whitney, N = 22, p = .0269; the difference between medium and high is weakly significant, N = 23, p = .0647.

13 Once more, on the mirror model that ignores censoring and clustering, the Hausman test turns out significant, but the problem can be remedied by a Hausman Taylor model.
already shown that punishment has a positive effect on changes of contributions (Table 5), we can conclude that the reaction to punishment is imperfect. Punishment induces participants to adjust their behavior. But they do not immediately realign with the implicit group norm.

In treatment high, this is patent. All participants know how much all other individual participants have contributed in each period, and by how much they have been punished in response. In treatment medium, participants only have a noisy signal of the sensitivity to punishment. In treatment low, they only see two things: how they have been punished themselves, and the effect of total punishment on contributions. Interestingly, since we have shown that contribution dynamics do not differ across treatments for those who have not been punished (Table 5), the differences in information across treatments are not relevant for these participants. Primarily, they seem to orient themselves to the contribution level in the previous period (Table 4). By contrast, those who have been punished are sensitive to this information (Table 5).

| received punishment points in previous period | -.849* |
| Low | 8.419** |
| Lrecpun*low | .561 |
| medium | 7.143* |
| Lrecpun*medium | .018 |
| final period | -2.097* |
| Cons | 17.941*** |
| N | 1296 |
| left censored | 48 |
| right censored | 723 |
| p model | <.001 |

Table 6
Effect of Received Punishment on Absolute Contributions

The critical question is this: on which pieces of information do participants condition their behavior – in particular those who have been punished in the previous period? Since treatments mainly differ with respect to contribution dynamics, in the following we focus on this dependent variable.

In all treatments, two pieces of information are available. The participant recalls how severely and how consistently she has been punished herself. We measure severity the following way: for each participant and period, we calculate the ratio of received punishment relative to the tokens not invested:

\[ \text{indpunrat} = \frac{\text{recpun}}{20 - \text{contr}} \]
Our first independent variable is the average of this measure for all past periods, per individual. The second independent variable measures the consistency of punishment as the standard deviation of the severity variable, again for all past periods and separately per individual. We also interact these two variables with received punishment in the previous period. The main effect for past severity is insignificant, but there is a highly significant interaction with received punishment. The more punishment has been severe in the past, the less punishment now is effective (models 1 and 3).\textsuperscript{14} There is also a weakly significant effect of past punishment consistency, in the expected direction: the more punishment has been inconsistent in the past, the less participants increase their contributions now (model 2). Interestingly, if one adds past severity and its interaction with punishment in the previous period, the main effect of past inconsistency is no longer significant. We then, however, establish a significant interaction with received punishment in the previous period (model 3). Surprisingly, the higher punishment now, the more participants increase their contributions if punishment has been (fairly) inconsistent in the past. This finding is, however, in line with research aiming at establishing a potentially beneficial effect of ambiguity in (criminal) law (Baker, Harel et al. 2004; Horovitz and Segal 2006). With these qualifications, we thus support H\textsubscript{2}.

<table>
<thead>
<tr>
<th></th>
<th>model 1</th>
<th>model 2</th>
<th>model 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>received punishment points in previous period</td>
<td>$1.636^{***}$</td>
<td>$.807^{***}$</td>
<td>$1.646^{***}$</td>
</tr>
<tr>
<td>past severity of individual punishment</td>
<td>$-1.260$</td>
<td>$-1.322$</td>
<td></td>
</tr>
<tr>
<td>L\textsubscript{recpun}*L\textsubscript{sevindpun}</td>
<td>$-1.460^{***}$</td>
<td>$-1.822^{***}$</td>
<td></td>
</tr>
<tr>
<td>past inconsistency of individual punishment</td>
<td>$-1.481^{*}$</td>
<td>$.922$ $^{**}$</td>
<td></td>
</tr>
<tr>
<td>L\textsubscript{recpun}*Lin\textsubscript{conindpun}</td>
<td>$-.319$</td>
<td>$.907^{**}$</td>
<td></td>
</tr>
<tr>
<td>Low</td>
<td>.653</td>
<td>.495</td>
<td>.535</td>
</tr>
<tr>
<td>Medium</td>
<td>.426</td>
<td>.446</td>
<td>.544</td>
</tr>
<tr>
<td>Final</td>
<td>$-1.342^{***}$</td>
<td>$-1.348^{***}$</td>
<td>$-1.328^{***}$</td>
</tr>
<tr>
<td>Cons</td>
<td>$-.701^{**}$</td>
<td>$.447$</td>
<td>$-.624^{*}$</td>
</tr>
<tr>
<td>N</td>
<td>1296</td>
<td>1150</td>
<td>1150</td>
</tr>
<tr>
<td>R\textsuperscript{2} within</td>
<td>.3062</td>
<td>.2645</td>
<td>.3332</td>
</tr>
<tr>
<td>R\textsuperscript{2} between</td>
<td>.0731</td>
<td>.0318</td>
<td>.0406</td>
</tr>
<tr>
<td>R\textsuperscript{2} overall</td>
<td>.2221</td>
<td>.1584</td>
<td>.2361</td>
</tr>
<tr>
<td>p model</td>
<td>&lt;.001</td>
<td>&lt;.001</td>
<td>&lt;.001</td>
</tr>
</tbody>
</table>

Table 7
Explaining Changes in Contributions with Individual Punishment Experiences

<table>
<thead>
<tr>
<th></th>
<th>model 1</th>
<th>model 2</th>
<th>model 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>received punishment points in previous period</td>
<td>$1.636^{***}$</td>
<td>$.807^{***}$</td>
<td>$1.646^{***}$</td>
</tr>
<tr>
<td>past severity of individual punishment</td>
<td>$-1.260$</td>
<td>$-1.322$</td>
<td></td>
</tr>
<tr>
<td>L\textsubscript{recpun}*L\textsubscript{sevindpun}</td>
<td>$-1.460^{***}$</td>
<td>$-1.822^{***}$</td>
<td></td>
</tr>
<tr>
<td>past inconsistency of individual punishment</td>
<td>$-1.481^{*}$</td>
<td>$.922$ $^{**}$</td>
<td></td>
</tr>
<tr>
<td>L\textsubscript{recpun}*Lin\textsubscript{conindpun}</td>
<td>$-.319$</td>
<td>$.907^{**}$</td>
<td></td>
</tr>
<tr>
<td>Low</td>
<td>.653</td>
<td>.495</td>
<td>.535</td>
</tr>
<tr>
<td>Medium</td>
<td>.426</td>
<td>.446</td>
<td>.544</td>
</tr>
<tr>
<td>Final</td>
<td>$-1.342^{***}$</td>
<td>$-1.348^{***}$</td>
<td>$-1.328^{***}$</td>
</tr>
<tr>
<td>Cons</td>
<td>$-.701^{**}$</td>
<td>$.447$</td>
<td>$-.624^{*}$</td>
</tr>
<tr>
<td>N</td>
<td>1296</td>
<td>1150</td>
<td>1150</td>
</tr>
<tr>
<td>R\textsuperscript{2} within</td>
<td>.3062</td>
<td>.2645</td>
<td>.3332</td>
</tr>
<tr>
<td>R\textsuperscript{2} between</td>
<td>.0731</td>
<td>.0318</td>
<td>.0406</td>
</tr>
<tr>
<td>R\textsuperscript{2} overall</td>
<td>.2221</td>
<td>.1584</td>
<td>.2361</td>
</tr>
<tr>
<td>p model</td>
<td>&lt;.001</td>
<td>&lt;.001</td>
<td>&lt;.001</td>
</tr>
</tbody>
</table>

\textsuperscript{14} On all mirror models that ignore clustering, the Hausman test turns out significant, but the problem can be remedied by a Hausman Taylor model. In models 2 and 3, the N is smaller since calculating the standard deviation is impossible for a single observation, per participant.
In treatments *medium* and *high*, participants also learn how intensely the supervisor has exercised her punishment power. Do they use this information? As model 1 in Table 8 shows, there is no significant main effect of the information about total punishment in the previous period. However the interaction effect is significant. Since we know that, overall, those unpunished do not change their contribution level, we can conclude that this information works as one would expect: those who have been punished react more sensitively if the supervisor punishes more intensely. This can be interpreted as a (small) severity effect, in line with H2.

<table>
<thead>
<tr>
<th></th>
<th>model 1</th>
<th>model 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>not punished in previous period</td>
<td>-2.188***</td>
<td>-2.609***</td>
</tr>
<tr>
<td>total punishment in previous period</td>
<td>.072</td>
<td></td>
</tr>
<tr>
<td>past sensitivity of punishment level to contribution level</td>
<td></td>
<td>.103</td>
</tr>
<tr>
<td>Linnoc*Ltotpun</td>
<td>-.171*</td>
<td></td>
</tr>
<tr>
<td>Linnoc*Lpunrat</td>
<td>-.343</td>
<td></td>
</tr>
<tr>
<td>Medium</td>
<td>.296</td>
<td>.304</td>
</tr>
<tr>
<td>Final period</td>
<td>-1.086*</td>
<td>-1.086*</td>
</tr>
<tr>
<td>cons</td>
<td>1.672***</td>
<td>2.021***</td>
</tr>
<tr>
<td>N</td>
<td>864</td>
<td>864</td>
</tr>
<tr>
<td>p model</td>
<td>&lt;.001</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>R² within</td>
<td>.1764</td>
<td>.1669</td>
</tr>
<tr>
<td>R² between</td>
<td>.1039</td>
<td>.0942</td>
</tr>
<tr>
<td>R² overall</td>
<td>.1351</td>
<td>.1287</td>
</tr>
</tbody>
</table>

Table 8
Effect of Additional Information Available in *medium* and *high*

*depvar: changes in contributions*
Random effects, robust standard errors, clustered for groups
Linnoc: not punished in previous period; Ltotpun: total punishment in previous period;
Lpunrat: past sensitivity of punishment level to contribution level

*** p < .001, ** p < .01, * p < .05

Since participants both know the development of contributions and of punishment, they are also in a position to estimate their supervisor’s sensitivity to the level of contributions. To measure the severity of punishment for contributions below the possible maximum contribution we define the variable *lavpunrat* which is calculated the following way: in a first step, for each group and period,

\[
punrat = \frac{\text{total punishment}}{20 - \text{average contribution}}
\]

is calculated. This is averaged over all past periods. The result is lagged by one period. As model 2 of Table 8 shows, there is no significant reaction to this information (p = .589).

---

15 Again the Hausman test turns out significant, but the problem can be remedied by a Hausman Taylor model.
In treatment high, participants see individual contributions and how the supervisor reacted to them. How does this additional information influence their behavior, in particular the behavior of those who have been punished in the previous period? Participants are indeed sensitive to information about the lowest contribution in the previous round (Table 9, model 1). More interestingly even, those who have been punished do not react the same way as those who have not. For those who have not been punished, we must add up the negative main effect and the positive interaction effect. Since the balance is positive, such participants react favorably. The higher the minimum contribution in the previous period, the more they increase their contributions in the subsequent period. The opposite is true for those who have been punished. They show a perverse reaction. The higher the minimum contribution in the previous period, the less they increase their contribution now, despite the fact that they have been punished.

In treatment high, participants also see how far the minimum and maximum contributions have been apart in the previous period. Model 2 of Table 9 demonstrates that they also condition their behavior on this information. Again those who have been punished use this information in a diametrically opposed way to how it is used by those not punished. If the distance was big, this induces those who have been punished to adjust contributions upwards, whereas those who have not been punished adjust contributions downwards. This makes sense. It describes a movement of convergence towards the group mean.

In treatment high, participants also have more information about the relationship between contributions and punishment. Regressor punishment severity is the coefficient of an OLS model that regresses received punishment on contribution. Regressor punishment inconsistency is the standard error of this coefficient. Both the coefficient and the standard error are calculated for all previous periods. For instance in period 4, they cover information from periods 1-3. We do not find a statistically significant effect of information on punishment severity, model 3. The effect of information on punishment inconsistency again differs between those who have been punished and those who have not, model 4. The significant interaction effect is best interpreted from the perspective of those who have been punished. For them, the sign is not swapped. Consequently, the less punishment was consistent in the past, the less they increase their contributions now. Interestingly, if we interact severity and consistency, both consistency and the interaction are significant, model 5. Severity has a negative effect if it is coupled with high inconsistency. Conversely, severity only has a beneficial effect if the punisher is predictable.

---

16 If we estimate the models of Table 9 as random effects models, the Hausman test turns out significant. Since all models only work with time variant regressors, there is no need for a Hausman Taylor model. We just use the fixed effects model.
Table 9
Effect of Information Only Available in high on Contribution Dynamics

<table>
<thead>
<tr>
<th></th>
<th>model 1</th>
<th>model 2</th>
<th>model 3</th>
<th>model 4</th>
<th>model 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>not punished in previous period</td>
<td>-4.850**</td>
<td>-1.295</td>
<td>-4.060***</td>
<td>-3.795**</td>
<td></td>
</tr>
<tr>
<td>lowest contribution in previous round</td>
<td>-.134*</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Linnoc*Lmincontr</td>
<td>.151*</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>distance between minimum and maximum contribution in previous round</td>
<td>.196**</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Linnoc*Ldistcontr</td>
<td>-.273***</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>past severity of punishment</td>
<td>-3.793</td>
<td>9.955</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Linnoc*sevpun</td>
<td>4.110</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>past inconsistency of punishment</td>
<td>-4.008</td>
<td>-49.079*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Linnoc*inconstpun</td>
<td>12.450</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sevpun*inconstpun</td>
<td>-162.030***</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>final period</td>
<td>-.002</td>
<td>-.184</td>
<td>.078</td>
<td>.106</td>
<td>-.020</td>
</tr>
<tr>
<td>cons</td>
<td>3.523***</td>
<td>.638</td>
<td>1.428</td>
<td>2.703***</td>
<td>5.508*</td>
</tr>
<tr>
<td>N</td>
<td>432</td>
<td>432</td>
<td>428</td>
<td>428</td>
<td>428</td>
</tr>
<tr>
<td>p model</td>
<td>&lt;.001</td>
<td>&lt;.001</td>
<td>.0007</td>
<td>&lt;.001</td>
<td>.0015</td>
</tr>
<tr>
<td>R² within</td>
<td>.1218</td>
<td>.1404</td>
<td>.1173</td>
<td>.1126</td>
<td>.1377</td>
</tr>
<tr>
<td>R² between</td>
<td>.1073</td>
<td>.1311</td>
<td>.1046</td>
<td>.1049</td>
<td>.0030</td>
</tr>
<tr>
<td>R² overall</td>
<td>.0876</td>
<td>.1047</td>
<td>.0838</td>
<td>.0825</td>
<td>.0422</td>
</tr>
</tbody>
</table>

f. Testing the Hypotheses

What do we learn from Phase 1 results for our hypotheses? In our experiment, letting everyone see the criminal system in action does not help, but hurts. We refute \( H_1 \). In line with \( H_2 \), participants are sensitive to the severity and consistency of punishment they have received themselves, Table 7. Information about average sensitivity of punishment to the average level of contributions is immaterial though, but those who have been punished in the previous period are more influenced by information about the total amount of punishment than those not punished, Table 8. Information about punishment consistency is only available in treatment high. In this treatment, contributions are lowest. This seems to speak against the relevance of overall consistency for deterrence. Evidence from treatment high points into the same direction. If consistency information is provided, only those who have been punished are sensitive to it, Table 9. The solution is provided by model 5 in Table 9. Consistency and severity interact. While severity does not have a beneficial main effect, it aggravates the negative effect of inconsistency. This qualifies \( H_2 \).
We partly support \( H_3 \). Those who have not been punished do indeed contribute more, Table 6. However, information about total punishment and about the average sensitivity of punishment to behavior is immaterial for this, Table 8, as is information about the relationship between behavior and punishment at the individual level, Table 5. For those not punished, the criminal system exclusively matters indirectly, through its effect on the general willingness to contribute to the common cause, Table 4.

3. Foreign Experiences

In period 11, i.e. in the first period of phase 2, participants contribute significantly more in treatment low versus high (Mann Whitney, \( N = 24 \), \( p = .0371 \)) and in treatment medium versus high (Mann Whitney, \( N = 24 \), \( p = .0485 \)), while the difference between low and medium is insignificant. Comparing contributions, per group, between periods 1 and 11, there is a significant difference in treatment low (Wilcoxon, \( N = 24 \), \( p = .0205 \)), but not in the remaining treatments. Hence the disadvantage of full transparency carries over to the next generation of players. Knowing everything about how strangers have been treated by the supervisor is similarly detrimental as knowing the same about one’s own community.

In period 11, new active players can only rely on foreign experiences to guide their behavior. They have not yet made their own experiences. As the regressions in Table 10 show, vicarious learning does indeed take place (Stafford and Warr 1993) (for the psychological background, see Bandura 1977). By far the most important factor, both in terms of significance and effect size, is the one that also had the highest explanatory power during phase 1: average contributions. But the development of contributions over time also has a significantly positive effect. This independent variable is the coefficient of a Tobit model that regresses contributions on period, controlling for the endgame effect by an additional coefficient for period 10, separately for each group.

Model 2 checks the effect of those additional pieces of information about the first phase that are only available in treatments medium and high. Actually, the additional information is of little relevance. The only weakly significant effect concerns severity of punishment. It has a strongly negative effect in treatment high, while it is beneficial in treatment medium. This independent variable is the coefficient of a linear regression explaining received punishment in the first phase with contributions, separately for each group. None of the additional pieces that are only available in high has explanatory power.
Table 10
Explaining Contributions in Period 11
Tobit, robust standard errors, clustered for groups
avcontrphase1: average contributions in phase 1
trendcontrphase1: development of contributions over time in phase 1
punphase1: average punishment in phase 1; punsev1: severity of punishment in phase 1
*** p < .001, ** p < .01, * p < .05, + p < .1

The most important piece of news concerns H3. Players are not only contributing more if they know that their own group has contributed more. It suffices if they learn that a group managed by the same supervisor has made high contributions, Figure 4. Vicarious experiences matter the same way as own experiences. Comparing treatments medium and high, we again refute H1. The effect comes through the fact that punishment information (both regarding the level of punishment and its severity) is counterproductive in treatment high, while it is productive in treatment medium, which also partially supports H2.
4. Own and Foreign Experiences Interacting

From period 12 on, own and foreign experiences interact.\textsuperscript{17} The two models of Table 11 contrast a static with a dynamic perspective. In both perspectives, the fundamental picture is the same as without foreign experiences. Absolute contributions are chiefly explained by average contributions in the previous round ($z = 4.482$). Note that the coefficient is even bigger than 1. In the following period, participants on average contribute even more than they had on average contributed in the previous period. In a dynamic perspective, again received punishment is the strongest predictor ($z = 13.21$). If they have been punished, participants increase their contributions – the more so, the more punishment they have received.

The level of contributions in the group of predecessors is insignificant in the model explaining levels, and it is only marginally significant in the model explaining first differences. Since this independent variable was crucial for explaining contributions in the initial period of the second group (Table 10), we conclude that vicarious experiences are replaced by own experiences as soon as they are available. In a dynamic perspective, the trend of contributions in the predecessor group keeps explanatory power. Yet through the interaction effects, it becomes clear that this only matters for treatment $\text{high}$. This makes sense. From the second period on, participants have their own experiences regarding the level of contributions, but they lack experiences about con-

\textsuperscript{17} Again, for all models in, the Hausman test is significant, but the problem can be remedied by a Hausman Taylor model.
tribution dynamics. In treatment high they use the information about the predecessor group as a proxy.

For the dynamic model, in treatment low information about the level of contributions in the predecessor group has a more positive effect than in the other treatments. Yet it remains that received punishment is less effective here than in the remaining two treatments.

<table>
<thead>
<tr>
<th></th>
<th>model 1 depvar absolute contributions</th>
<th>model 2 depvar first differences</th>
</tr>
</thead>
<tbody>
<tr>
<td>low</td>
<td>-10.558 *</td>
<td>-2.331*</td>
</tr>
<tr>
<td>medium</td>
<td>4.322</td>
<td>.128</td>
</tr>
<tr>
<td>average contribution per group in previous period</td>
<td>1.031***</td>
<td></td>
</tr>
<tr>
<td>lavcontr*low</td>
<td>.063</td>
<td></td>
</tr>
<tr>
<td>lavgrp*medium</td>
<td>-.258</td>
<td></td>
</tr>
<tr>
<td>received punishment in previous period</td>
<td>-.115</td>
<td>1.096***</td>
</tr>
<tr>
<td>lrecpun*low</td>
<td>-.153</td>
<td>-.311*</td>
</tr>
<tr>
<td>lrecpun*medium</td>
<td>-.750*</td>
<td>-.054</td>
</tr>
<tr>
<td>average contributions in phase 1</td>
<td>.108</td>
<td>-.076*</td>
</tr>
<tr>
<td>low*avcontrphase1</td>
<td>.723</td>
<td>.178*</td>
</tr>
<tr>
<td>medium*avcontrphase1</td>
<td>.186</td>
<td>.018</td>
</tr>
<tr>
<td>development of contributions over time in phase 1</td>
<td>2.596</td>
<td>.310**</td>
</tr>
<tr>
<td>low*trendcontrphase1</td>
<td>-2.443</td>
<td>-.308**</td>
</tr>
<tr>
<td>medium* trendcontrphase1</td>
<td>-2.085</td>
<td>-.282**</td>
</tr>
<tr>
<td>final period</td>
<td>-3.838***</td>
<td>-1.472***</td>
</tr>
<tr>
<td>cons</td>
<td>.409</td>
<td>.385</td>
</tr>
<tr>
<td>N</td>
<td>1296</td>
<td>1296</td>
</tr>
<tr>
<td>left censored</td>
<td>24</td>
<td></td>
</tr>
<tr>
<td>right censored</td>
<td>814</td>
<td></td>
</tr>
<tr>
<td>$R^2$ within</td>
<td></td>
<td>.3252</td>
</tr>
<tr>
<td>$R^2$ between</td>
<td></td>
<td>.1601</td>
</tr>
<tr>
<td>$R^2$ overall</td>
<td></td>
<td>.2581</td>
</tr>
<tr>
<td>p model</td>
<td>&lt;.001</td>
<td>&lt;.001</td>
</tr>
</tbody>
</table>

Table 11
Explaining Contributions and Contribution Changes in Periods 12-20
model 1: random effects Tobit, bootstrapped at the group level, 500 reps
model 2: random effects, robust standard errors, clustered for groups
lavcontr: average contribution per group in previous period
lrecpun: received punishment in previous period; avcontrphase1: average contributions in phase 1
trendcontrphase1: development of contributions over time in phase 1

*** p < .001, ** p < .01, * p < .05, * p < .1
VI. Conclusions

Experiments are not meant to portray the richness of the criminal system. They deliberately strip off context, in order to establish full control over the one feature that purportedly makes criminal sanctions effective. Jeremy Bentham had the conjecture that punishment must be made transparent if it is to guide those who have not been punished themselves.

Given our results, we cannot but conclude: as a policy maker, Jeremy Bentham did not seem to get it right. At least in our experiment, if contributions and punishments are transparent, the willingness to contribute to the common cause decays. Even more disturbingly: actual sanctions become less effective, the better would-be low contributors can observe how (mildly) the criminal system reacts to offenses. The results suggest that instead of building panopticon prisons in the town centre, government should conceal prisons from public scrutiny. Punishment serves society best if it remains a tool one does not see in action when applied to others. What really matters is information about normabiding behavior of other members of society. To use a metaphor that features prominently in criminal policy: government should spend money on repairing broken windows, not on showcasing correctional action.

Our experimental result is not only at variance with Jeremy Bentham’s advice. It also qualifies a strand of the criminological literature. We do find the expected effects of punishment severity and consistency if we consider the sanctions this one individual has received in the past. In the field, the effect might, however, be blurred by the fact that the second offense will often be qualitatively different from the first. Yet normally, if one speaks of consistency and severity as policy variables, one rather thinks of them at the level of society. At any rate, greater transparency only has an effect on the latter. In our setting overall punishment severity occasionally has a minor effect, as has punishment consistency. But both effects are at most marginally significant. What looks like the prime decision variable of criminal policy turns out to be at best secondary, while the main interest should be in the effects of punishment on reducing the amount and severity of crime in the first place.

Nonetheless, this is a paper in the spirit of Jeremy Bentham. While he might have got it wrong as a policy maker, he got it totally right as an analyst. The main task of criminal policy is and ought to be that would-be criminals are induced to exhibit socially desirable behavior. Only the route to the end is a different one. The main tool ought to be impression management, not deterrence.
References


Appendix A

Individual Groups

Figure 5
Contributions and Punishment per Group
**Box 1
Strategy for Parametric Estimation**

We have data from periods, nested in subjects, nested in groups. In many periods, many subjects have contributed their entire endowment of 20. In some periods, some subjects have kept their entire endowment for themselves. Consequently, the contribution data is heavily right censored and slightly left censored. There are standard procedures to separately deal with both features of the data. The nested character is captured by a random or fixed effects model with standard errors clustered at the group level. Censoring is taken into account in a (random effects) Tobit model. Unfortunately there is no standard procedure for clustered and censored data. We solve the problem by bootstrapping the ordinary random effects model, with random draws clustered at the group level. Note that, if one were to ignore the fact that observations are not independent for members of one and the same group, i.e. if one did not cluster standard errors, coefficients would still be unbiased. Only the standard errors might be too small. Bootstrapping precisely targets this problem. It generates standard errors that take the clustering into account. In these estimations, we always work with 500 independent draws.

A second statistical problem results from the fact that there is no generally acknowledged fixed effects estimator for censored data. Therefore we cannot perform the Hausman test on the random effects Tobit model. To nonetheless make sure that we are justified to use the more efficient random effects model, we rerun a mirror model that ignores censoring. This of course yields a standard random effects model. We rerun the model as a fixed effects model, and compare the coefficients by way of a Hausman test. In some models, the Hausman test turns out significant. We then are able to remedy the problem by a Hausman Taylor model. This saves the time invariant regressors, in particular the treatment effects, which would cancel out in the fixed effects model via mean differencing. The Hausman Taylor model instruments the time variant regressors by the deviation from their mean, as in the fixed effects model, and it instruments the time invariant regressors by the mean of the exogenous time variant regressors (Hausman and Taylor 1981). A further Hausman test demonstrates that, in all instances where we need this procedure, it has been successful (Baltagi, Bresson et al. 2003).
**Instructions for Treatment High Second Phase**

(instructions for phase 1 and for phase 2, treatments low and medium are available upon request)

**General Instructions for Participants**

You are about to take part in an economics experiment. If you read the following instructions carefully, you will be able to earn a substantial sum of money, depending on the decisions you make. It is therefore very important that you read these instructions carefully.

The instructions you have received are exclusively for your private information. **There shall be absolutely no communication during the experiment.** If you have questions, please ask us. Disobeying this rule will lead to exclusion from the experiment and any payments.

The experiment consists of several parts. We will begin by explaining the first part. You will receive separate instructions for the other parts.

You will definitely receive €2.50 for participating in the experiment. During the experiment, the currency in operation is not euro, but taler. Your entire income is hence first calculated in taler. The total number of taler you will have accumulated in the course of the experiment will then be transferred into euro at the following rate:

$$1 \text{ taler} = 3 \text{ euro cent}.$$  

At the end of the experiment, you will receive a **cash** payment, in euro, of whatever number of taler you have earned.

Participants are divided into groups of five. In other words, there are 4 further participants in your group.

All five participants in your group are taking part in this experiment for the first time. There are two roles: four participants, who have confirmed their presence at this experiment for two hours, are assigned Role A. Another person, who has confirmed his presence at this experiment for 4 hours, is assigned Role B.

The experiment is divided into individual periods, of which there are a total of 10. During these 10 periods, the constellation of your group of five remains unchanged. **You are therefore in the same group with the same people for 10 periods. During these 10 periods, the role you have been assigned also remains unchanged.**

At the beginning of the experiment, each participant is given a lump-sum payment of 50 taler. This occurs only once. You may cover possible losses with these 50 taler.
The following pages give you an outline of the exact proceedings of the experiment.

Information on the Exact Proceedings of the Experiment

Each of the 10 periods consists of two steps. In Step 1, the participants who have been assigned Role A decide on contributions to a project. In Step 2, the participant who has Role B can reduce the income of the other (Role A) participants. At the beginning of each period, each participant receives 20 points, referred to henceforth as endowment.

**Step 1:**

In Step 1, only the four Role A participants in a group make a decision (should you have been assigned Role B, please read this part of the instructions anyway, in order to find out how a Role A participant can reach a decision). Your task is to reach a decision on how to use your endowment. As a Role A participant, you have to decide how many of the 20 points you wish to pay into a project, and how many you wish to keep for yourself. The consequences of this decision are explained in more detail below.

At the beginning of each period, the following input screen appears:

**The input screen:**

![Input screen for Step 1](image-url)
On the top left corner of your screen, the period number is displayed.

As already mentioned, your endowment in each period is 20 points. As a Role A participant, you have to make a decision on your project contribution by typing in a sum between 0 and 20 in the appropriate field. You may click on this window by using the mouse. As soon as you have determined the sum you wish to contribute, you have also decided on how many points you keep for yourself: **20 minus your contribution**. Once you have keyed in your amount, press or click **O.K.**, using the mouse or the Enter-key. As soon as you have done this, you can no longer make any changes to your decision.

Your **income** from the contribution phase consists of two parts:

1. the points that you have kept for yourself ("income from endowment kept")
2. the "income from the project". The income from the project is calculated as follows:

\[
\text{Your income from the project} = 0.4 \times \text{the total sum of contributions to the project}
\]

Your income from the project, in taler, for one period is therefore

\[(20 \text{ minus your contribution to the project}) + 0.4 \times \text{(total sum of contributions to the project)}\].

The income of all other group members is calculated according to the same formula, i.e., each group member receives the same income from the project. If, for example, the sum of contributions from all group members is 60 points, then you and all other group members will receive a points income from the project of \(0.4 \times 60 = 24\) taler. If the group members have contributed a total of 9 points to the project, you and all other group members will receive \(0.4 \times 9 = 3.6\) taler as your income from the project.

For each point that you keep for yourself, you earn an income of 1 taler. If instead you contribute one point from your endowment to your group project, then the sum of contributions to the project increases by 1 point, and your income from the project increases by \(0.4 \times 1 = 0.4\) taler. However, this also means that the income of all other group members increases by 0.4 taler, so that the total income of the group increases by \(0.4 \times 5 = 2\) taler. Through your contributions to the project, the other group members also increase their earnings. On the other hand, you also earn something from the contributions of the other group members to the project. For each point that another group member contributes to the project, you earn \(0.4 \times 1 = 0.4\) taler.

Please be aware that the Role B participant in a group cannot contribute to the project. This participant receives the same income from the project as each Role A participant.
Step 2:

In Step 2, only the Role B participant in each group decides (should you have been assigned Role A, please read this part of the instructions anyway, in order to find out how a Role B participant can reach a decision). As a Role B participant, you can reduce or leave unchanged the income of each of the other participants in Step 2, namely by assigning “points”. This becomes clear once you take a look at the input screen for Step 2:

The input screen for Step 2

Here you can see how much the individual Role A group members have contributed to the project in this period. Please bear in mind that in each period the order in which the members of your group are displayed remains the same. Group members can be identified from period to period.

Now it is up to you to decide for each Role A group member in this period whether you wish to allocate points and how many points you wish to distribute. Whatever you decide, you are obliged to enter a figure. If you do not wish to change the income of one particular group member, please enter 0. If you enter a number higher than 0, you reduce this group member’s income.
You can move within the input fields under the heading “points” by using the tabulator key (→) or the mouse.

If you allocate points, this costs you taler; the amount depends on the number of points you allocate. Points are **whole numbers between 0 and 20**. The more points you allocate to a member of your group, the higher your costs are. The following formula gives you the correlation between points allocated and the costs of this allocation in taler:

VII. **Cost of points allocated = Number of points allocated.**

Each allocated “point” therefore costs you 1 taler. For instance, if you allocate 2 points to a member, this costs you 2 taler; if, in addition, you allocate 9 points to another group member, this costs you 9 taler; if you allocate 0 points to the two other group members, there is no cost. You have therefore allocated a total of 11 points and your **total cost** is, hence, 11 taler (2+9+0+0). If you press **Kostenberechnung (Calculate cost)**, the total cost is shown to you. Unless you have already clicked **Continue**, you may still change your decision.

If you choose 0 points for a particular group member, you do not change this group member’s income. If, however, you allocate **one** point to a member (i.e., if you choose 1), you **reduce** this member’s income by **3 taler**. If you allocate **2** points to a group member (i.e., if you choose 2), you reduce this member’s income by **6 taler**, etc. **For each point that you allocate to another group member, this member’s income is reduced by 3 taler.**

Please be aware that the Role A participants in a group cannot allocate points. The participant who has been assigned Role B can therefore not receive points in Step 2.

The total taler income of a Role A participant after both steps is hence calculated according to the following formula:

\[
\text{Taler income at the end of Step 2 for Role A} = \text{Period income for Role A} = \\
= \text{Income from Step 1} - 3 \times (\text{sum of points received})
\]

The total taler income of a Role B participant is hence calculated according to the following formula after both steps:

\[
\text{Taler income at the end of Step 2 for Role B} = \text{Period income for Role B} = \\
= \text{Income from Step 1} - \text{Cost of points allocated by you}
\]
Please bear in mind that the taler income can also be negative for Role A participants at the end of Step 2. This could be the case whenever the income reduction from points received is higher than the income from Step 1.

Once all participants have made their decision, a screen informs you of your period income and total income thus far.

**The income screen at the end of Step 2:**

If you have been allocated Role A, your screen looks as follows:
If you have been allocated Role B, your screen looks as follows:

Your total income at the end of the experiment is the sum of the period incomes according to the following formula:

\[
\text{Total income (in taler) from the experiment} = 50 + \text{Sum of all period incomes, if the total is not negative.}
\]

Otherwise, you receive 0 taler

In addition, you are given the sum of 2.50 euro for showing up.

As mentioned above, the member of your group who has been assigned Role B will take part in the same experiment on a further occasion. At the beginning of this future experiment, the new Role A participants, who will then form a group together with your participant B, will receive a chart for their information. This chart depicts the average contributions and the individual contributions as well as the points received by the four individual Role A participants from your cur-
rent group over 10 periods. The four Role A participants from the future experiment will be different participants to those in this experiment. Only the Role B participant is the same person. The participants in the new group will be told that the chart depicts the behavior of the former group with the same Role B participant.}

Do you have any further questions? If you do, please raise your hand from your booth – one of the experiment supervisors will be with you shortly.
Preprints 2010

2010/05: Gropp R., Hakenes H., Schnabel I., Competition, Risk-Shifting, and Public Bail-out Policies

2010/04: Slemrod J., Traxler C., Optimal observability in a linear income tax

2010/03: Baumann, Florian; Friehe, Tim; Grechenig, Kristoffel: Switching Consumers and Product Liability: On the Optimality of Incomplete Strict Liability

2010/02: Bierbrauer F., Hellwig M., Public-Good Provision in a Large Economy

2010/01: Bierbrauer F., Incomplete contracts and excludable public goods

Preprints 2009

2009/39: Petersen N., The Reception of International Law by Constitutional Courts through the Prism of Legitimacy


2009/37: Engel C., Das schwindende Vertrauen in die Marktwirtschaft und die Folgen für das Recht


2009/33: Hakenes H., Schnabel I., Credit Risk Transfer and Bank Competition

2009/32: Jansen J., Beyond the Need to Boast: Cost Concealment Incentives and Exit in Cournot Duopoly


2009/29: Engel C., Rockenbach B., We Are Not Alone: The Impact of Externalities on Public Good Provision


2009/27: Hahmeier M., Prices versus Quantities in Electricity Generation

2009/26: Burhop C., The Transfer of Patents in Imperial Germany

2009/25: Burhop C., Lübbers T., The Historical Market for Technology Licenses: Chemicals, Pharmaceuticals, and Electrical Engineering in Imperial Germany

2009/24: Engel C., Competition as a Socially Desirable Dilemma Theory vs. Experimental Evidence


2009/22: Traxler C., Majority Voting and the Welfare Implications of Tax Avoidance


2009/19: Magen S., Rechtliche und ökonomische Rationalität im Emissionshandelsrecht
2009/18: Broadberry S.N., Burhop C., Real Wages and Labour Productivity in Britain and Germany, 1871-1938: A Unified Approach to the International Comparison of Living Standards

2009/17: Glöckner A., Hodges S.D., Parallel Constraint Satisfaction in Memory-Based Decisions

2009/16: Petersen N., Review Essay: How Rational is International Law?


2009/15: Bierbrauer F., On the legitimacy of coercion for the financing of public goods

2009/14: Feri F., Irlenbusch B., Sutter M., Efficiency Gains from Team-Based Coordination – Large-Scale Experimental Evidence


2009/12: Hellwig M., Utilitarian Mechanism Design for an Excludable Public Good


2009/11: Weinschenk P., Persistence of Monopoly and Research Specialization

2009/10: Horstmann N., Ahlgrimm A., Glöckner A., How Distinct are Intuition and Deliberation? An Eye-Tracking Analysis of Instruction-Induced Decision Modes


A Public-Goods Experiment with a Super-Additive Player


2009/07: von Weizsäcker C., Asymmetrie der Märkte und Wettbewerbsfreiheit

2009/06: Jansen J., Strategic Information Disclosure and Competition for an Imperfectly Protected Innovation


2009/04: Rincke J., Traxler C., Deterrence Through Word of Mouth

2009/03: Traxler C., Winter J., Survey Evidence on Conditional Norm Enforcement


2009/01: Beckenkamp M., Environmental dilemmas revisited: structural consequences from the angle of institutional ergonomics, issue 2009/01