Discussion Papers
Department of Economics
University of Copenhagen

No. 14-27

Judicial Error and Cooperation

Thomas Markussen, Louis Putterman and Jean-Robert Tyran

Øster Farimagsgade 5, Building 26, DK-1353 Copenhagen K., Denmark
Tel.: +45 35 32 30 01 – Fax: +45 35 32 30 00
http://www.econ.ku.dk

ISSN: 1601-2461 (E)

Judicial Error and Cooperation

Thomas Markussen, Louis Putterman and Jean-Robert Tyran*

November 14, 2014

Abstract

Cooperation can be induced by an authority with the power to mete out sanctions for free riders, but law enforcement is prone to error. This paper experimentally analyzes preferences for and consequences of errors in formal sanctions against free riders in a public goods game. With type I errors, even full contributors to the public good may be punished. With type II errors, free riders may go unpunished. We find that judicial error undermines cooperation and that the effects of type I and II errors are symmetric. To investigate their relative (dis-)like for error, we let subjects choose what type of error to prevent. By use of an incentive-compatible mechanism, we find that subjects prefer type II over type I errors. We find that the strength of this preference is fully in line with a motive to maximize income and does not indicate any additional psychological or fairness bias against type I errors.

Keywords: Public goods, sanctions, type I errors, type II errors, willingness to pay

JEL codes: H41, K4, C92

* Markussen: Department of Economics, University of Copenhagen, Denmark. Thomas.Markussen@econ.ku.dk ;
Putterman: Department of Economics, Brown University, Providence, RI, USA. Louis_Putterman@brown.edu ;
Tyran: Department of Economics, University of Vienna, Austria and Department of Economics, University of Copenhagen, Denmark, and CEPR (London). Jean-Robert.Tyran@univie.ac.at
We thank Matteo Rizzolli and participants at seminars in Aarhus, Copenhagen, Lund and Vienna for helpful comments and the Austrian Science Fund (FWF) under project I2027-G16.

1. Introduction

A large experimental literature investigates the capacity of sanctions to promote cooperation in social dilemmas. Most papers focus on informal, decentralized sanctions (e.g. Fehr and Gächter 2000). Recently, a number of papers focusing on formal, centralized or pooled sanctions have also emerged (e.g. Tyran and Feld 2006, O'Gorman, Henrich and Vugt 2009, Putterman, Tyran and Kamei 2011, Traulsen, Röhl and Milinski 2012, Zhang et al. 2012, Andreoni and Gee 2012, Grechenig, Nicklisch and Thöni 2013, Markussen, Putterman and Tyran 2014, Kamei, Putterman and Tyran, forthcoming). Results from these experiments generally show that sanctions are effective in terms of increasing pro-social behavior, such as contributions to a public good. Whether or not earnings are increased depends on the cost of the sanctions, the gains from cooperation, and whether some cooperation is also sustained in the absence of sanctions.

In the case of informal institutions, considerable attention has been devoted to the issue of mis-targeted sanctions, referred to as "perverse" or "antisocial" punishment, where people are sanctioned in spite of choosing pro-social actions (e.g. Cinyabuguma, Page and Putterman 2006, Herrmann, Thöni and Gächter 2008). The experimental literature on formal sanctions has focused much less on targeting. Essentially, it is assumed that formal sanctions are always well-targeted.[1] In fact, formal sanction systems, such as those administered by the state, are not free of error. Judicial error can loom in two basic ways: Sometimes innocent people are punished (type I error) and sometimes those guilty of wrong-doing are not (type II error).[2]

This paper studies in a laboratory experiment how both types of judicial error affect contributions to a public good when free riding is subject to formal sanctions. We exogenously vary two fundamental properties of the sanctioning environment, namely the probability of error and the severity of sanctions. Furthermore, while previous experimental papers on judicial error have mostly considered exogenous

[1] An important distinction is made in the literature concerning the role of the central authority. In some papers it is fully automated (i.e. written into the experimental software, as in e.g. Yamagishi 1986) in others it is played by an experimental subject (e.g. Gürerk, Irlenbusch and Rockenbach 2009, Heijden, Potters and Sefton 2009, Nosenzo and Sefton 2014 and one treatment in Carpenter, Kariv and Schotter 2012). In the latter type of experiment, mis-targeted sanctions may occur, but their nature and frequency cannot be controlled by the experimenter.

[2] There is no strong convention on this way of using the terms "type I" and "type II" error. For example, Harris (1970), Png (1986) and Polinsky and Shavell (2000) use the opposite definition of the one stated here. However, the papers most closely related to ours (e.g. Rizzolli and Stanca 2012 and Dickson, Gordon and Huber 2009) use the terms in the same way as we do. As noted by Dickson et al. 2009, one argument in favor of our definition is that it parallels the terms type I- and type II error in statistics. In legal matters, the "null hypothesis" is that the defendant is innocent. A type I error, by our definition, is the rejection of a true null-hypothesis (conviction of an innocent), while a type II error is the failure to reject a false null-hypothesis (not convicting a guilty defendant).

errors, we investigate people's preferences for type I versus type II errors by letting them choose which type of error to prevent.

Understanding preferences and consequences for types of judicial errors is important for the design of formal sanctioning systems. First, reducing errors is costly. Some types of expenditure, such as increased spending on police investigation, are likely to prevent both type I and type II errors. Other types of spending may have differential effects on the two types of error. For example, improving access to legal assistance for defendants is likely to reduce type I errors, while increased resources for public prosecutors may reduce type II errors. Second, the choice of legal standards and investigation procedures sometimes implies a trade-off between type I and type II errors. A clear example is the "burden of evidence". To obtain a legal conviction, the defendant must be proven guilty according to some "evidence standard", with a prime example being that the prosecution must provide inculpatory evidence "beyond reasonable doubt". Reducing the evidence standard increases the risk of type I errors and decreases the likelihood of type II errors. Also, some types of police operations (e.g. "stop and search" based on racial profiling) may be seen as reducing the risk of type I errors but increasing type II errors, to the extent that being profiled and searched is a punishment in itself.

If reduction of judicial error is costly, and the reduction of one type of error comes at the expense of the other, then it is important to understand i) how much and when judicial error affects behavior and ii) what are people's preferences for one type of error versus the other. The second issue is particularly relevant in democratic societies. It may well be that the two types of error have similar effects on preventing crime but that the population has a preference for preventing one type of error over the other. In that case, economic policy will not only be guided by considerations of the cost of reducing errors and effective deterrence but also by voters' preferences, e.g. their aversion to type I errors (punishing innocent people). But to know whether that it is the case, one needs a measure of aversion. Our paper sets out to contribute to both of the questions above, i.e. to the behavioral consequences of (one vs. the other type of) error and to the people's relative aversion against different types of judicial error.

Experimental methods are particularly apt to investigate both of these questions. The ability to know payoffs, error probabilities and information conditions and to vary them in a controlled way is a key advantage of experiments. In addition, we implement an incentive-compatible mechanism for eliciting preferences of preventing one type of error rather than the other. We investigate the issue in the context of a public goods game because we have situations in mind in which crime and its prevention have an externality dimension, i.e. affect a whole group of people. Property crime, like theft or burglary, obviously involves redistribution from the victim to the perpetrator and it involves efficiency costs because the

perpetrator expends effort and the victim perhaps expends effort to prevent the theft. In addition, such action may induce others (who have not or not yet been victimized) to make preventive efforts. For example, they may secure their houses with locks and alarms or, to take an extreme example, they may not dare any longer to leave their houses. The environment we use to investigate the issue (the public goods game) is admittedly a very stylized representation of that idea but has advantages because it is a well-established workhorse in the experimental literature. In addition, the aversion against judicial errors and the consequences of type I and type II errors have not been studied previously in a unified framework in a cooperation context, to the best of our knowledge.

Our main results are as follows.

R1: Sanctions that are free of judicial error improve cooperation. We find the usual inefficient under-provision of public goods absent sanctions, and confirm results from previous studies that error-free sanctions improve cooperation.

R2: Judicial error is harmful. We find that errors reduce contributions to the public good, i.e. they undermine the behavioral effects of sanctions. This reduction is significant when predicted by standard theory (i.e. when errors turn theoretically deterrent sanctions into non-deterrent ones), but we also find significant effects when such a deterrence-undermining effect is not predicted (i.e. when errors merely weaken an already non-deterrent sanction). The harmful effect of judicial errors is stronger for deterrent than for non-deterrent sanctions.

R3: The harm done by the two types of judicial error is the same. In line with the predictions of standard theory, we find that the behavioral effects of errors of type I and type II on contributions to the public good are symmetric. That is, an increase in type I error is equally harmful in terms of undermining incentives to cooperate as is an increase in type II error, all else equal.

R4: Judicial error is more harmful when the error probability is high.

R5: Subjects dislike errors of type I more than errors of type II. The willingness to pay to prevent type I error is higher than the willingness to prevent type II error.

R6: The strength of the relative dislike of the two types of judicial errors is fully explained by a motive to maximize income. We find no evidence of an additional dislike for type I errors based on risk aversion or social preferences.

Related literature

Dickson, Gordon and Huber (DGH, 2009) is in some ways the closest match to our paper. DGH experimentally study a public goods game with punishment and with type I and type II errors, as we do. However, there are a number of differences in design and in practical implementation which make a comparison of results difficult. Most importantly, punishment in DGH is administered by a subject in the experiment (denoted "monitor"), whereas in ours it is administered by the experimenter.[3] In DGH, monitors receive information about the contribution behavior of other subjects. This information is subject to error. Because actual punishment decisions are chosen by monitors, type I and type II errors are not exogenous. This contrasts with our experiment, where the probability of error is, at least initially, entirely exogenous. DGH find that the effects of type I and type II errors are at first indistinguishable but diverge after some periods of repeated play, with contributions declining faster under type I error than under type II error. This result is driven by the fact that monitors punish less often under type I than under type II errors. The effects of actual type I and type II errors on contributions in subsequent periods are symmetric. DGH include a treatment with endogenous choice of error regime, as we do, but in DGH it is the monitor, rather than the participants in the public goods game, who chooses between a system with type I errors and an error-free system. A system with type I error can be chosen at no cost while the error-free system is costly. DGH find that type I errors have a stronger, negative effect on contributions when they are chosen by the monitor than when they are imposed, although monitors make up for this effect by applying punishment more frequently in the treatment with endogenous error regimes.

Rizzolli and Stanca (2012) study the effect of type I and type II errors on pro-social behavior in a "theft game" - i.e. a reverse dictator game, where one subject may "steal" some of the endowment of another subject. They consider exogenous errors only. The authors find that type I errors have a slightly stronger effect on stealing behavior than type II errors. This difference is explained as an effect of risk aversion.

Sonnemans and van Dijk (2011) present an experiment which parallels the sentencing decisions of judges in the face of uncertain evidence. Results show that subjects frequently reach incorrect verdicts and that decisions are biased in the direction of type I errors (called "unfounded convictions"). The authors do not interpret this as indicating a preference for type I over type II errors, but rather as the result of systematic deviations from rationality in decision making.

A group of experimental papers study the effect of imperfect information in public goods games with informal, peer-to-peer sanctioning. In these experiments, subjects receive information about other

[3] Other differences include the use of binary contribution and punishment decisions in DGH vs. a range of integer contribution and sanction values to choose from in our experiment, and the use of partner matching in DGH vs. stranger matching in ours.

subjects' contributions to the public good, but this information may be incorrect. Therefore, it might result in type I or type II errors in informal punishment. Carpenter (2007), Grechenig, Nicklisch and Thöni (2010), and Ambrus and Greiner (2012) all find that peer-to-peer punishment combined with imperfect (noisy) or incomplete information significantly reduces contributions relative to the case with perfect information and often leads to significantly lower payoff than with no punishment option at all.

Various experimental papers have tested the deterrent effects of sanctions (e.g. Harbaugh, Mocan and Visser 2011, Schildberg-Hörisch and Strassmair 2012, Khadjavi 2014). There is also a number of experimental papers where type II errors play a role, while type I errors are ignored. For example, DeAngelo and Charness (2012) implement an experiment where participants choose between a profitable but illegal activity (roadway speeding) and a less profitable, legal one (not speeding). Speeding is detected with some probability lower than one, which means that type II errors occur. Results show, among other things, that the combination of severe punishment and low probability of detection (i.e. high probability of type II error) has a stronger deterrent effect than the payoff equivalent combination of milder punishment and higher probability of detection. Dai, Hogarth and Villeval (2014) study a public goods game with centralized punishment, where contributions are "audited" (and free riders consequently punished) with some probability. This probability is unknown to subjects. The paper focuses on analyzing what happens when audits are removed and results show that the effect of audits on cooperation lingers much longer when audits are infrequent and irregular than when they are frequent and regular. The reason is that it takes longer for subjects to learn that audits have been removed in the former case. There is also a related experimental literature on tax evasion (e.g. Alm, McClelland and Schulze 1992, Tan and Yim 2014), where tax evaders are caught with some probability lower than one, and an experimental literature on corruption, where acts of corruption are typically detected with some probability (Abbink 2006).

The paper is organized as follows. Section 2 presents theoretical considerations and derives predictions, Section 3 presents the experimental design and procedures. Section 4 presents results, and Section 5 concludes.

2. Theoretical considerations

This section explains the incentive effects that arise from type I and II error in the context of providing public goods as studied here both assuming rationality and self-interest and under alternative assumptions (that e.g. incorporate social preferences). We show that the behavioral effects of type I and type II error are symmetric in their capacity to undermine cooperation. We discuss our design choices from a theoretical

perspective and discuss the incentive-compatible mechanism. We also explain that rational and self-interested agents have a preference for avoiding type II over type I error, and we discuss under which conditions a more pronounced aversion can be rationalized.

We consider a society of n members facing a collective action problem. Each member i receives an endowment W and decides how much, c_i, of this endowment to allocate to the production of a public good and how much to keep for themselves. Subject i's payoff function is:

$$\pi_i = W - c_i + m\sum_{j=1}^{n} c_j \ , \tag{1}$$

where $1 > m > 1/n$, so that allocation to the public good is collectively but not individually optimal. In this standard, linear public goods environment, a system to sanction free riders is introduced. In particular, individuals pay a fine of s points for each point they allocate to their private account:

$$\pi_i = \left(W - c_i\right)\left(1 - s\right) + m\sum_{j=1}^{n} c_j \ . \tag{2}$$

When $s \geq 1 - m$, the sanction is *deterrent* in the sense that contributing to the public good is individually optimal (i.e. free riding is deterred). When $s < 1 - m$, we say that the sanction is "non-deterrent".

Table 1: Probability of errors

	Type I error	Type II error
Condition A	*P*	0
Condition B	0	*p*

Note: $0 < p < 1$.

We let this sanctioning system be prone to errors. With **type I error**, there is some risk that a fine of sW is imposed regardless of how the subject allocated the points. So, sometimes even cooperators (the "innocent") are punished. With **type II error**, there is some chance that no fine is imposed, regardless of how the subject allocated the points. So, sometimes even free riders (the "guilty") escape when type II errors prevail. In our baseline conditions, we study environments in which either type I or type II (but not both) errors occur with some exogenously determined probability p (see table 1).

The payoff functions (in expected terms) in Conditions A and B are, respectively:

Condition A (type I errors may occur):

$$\begin{aligned}\pi_i^A &= p\left(W(1-s)-c_i+m\sum_{j=1}^{n}c_j\right)+(1-p)\left((W-c_i)(1-s)+m\sum_{j=1}^{n}c_j\right)\\ &=(1-s)W+\left(m+(1-p)s-1\right)c_i+m\sum_{j\neq i}c_j\end{aligned} \tag{3}$$

Condition B (type II errors may occur):

$$\begin{aligned}\pi_i^B &= p\left(W-c_i+m\sum_{j=1}^{n}c_j\right)+(1-p)\left((W-c_i)(1-s)+m\sum_{j=1}^{n}c_j\right)\\ &=\left(1-(1-p)s\right)W+\left(m+(1-p)s-1\right)c_i+m\sum_{j\neq i}c_j\end{aligned} \tag{4}$$

Two properties of equations (3) and (4) are important to note. First, the coefficients on c_i are the same. Therefore, for given error probabilities *p*, the effect of type I and type II errors on the marginal payoff incentive to contribute to the public good *are the same*. Essentially, deterrence stems from the expected difference between payoff from being innocent and payoff from being guilty. Type I errors reduce the expected payoff from innocence, while type II errors increase the payoff from being guilty. Hence, the effects of the two types of error are symmetric as they both lower the difference in returns between the innocent and the guilty (Png 1986, Garoupa 1997, Polinsky and Shavell 2000).

Second, for identical error probabilities and contribution profiles (i.e. assuming that different types of error do not affect contributions differently),

$$\pi_i^B-\pi_i^A=psW>0\,. \tag{5}$$

Equation (5) says that, all else equal, expected earnings are *psW* higher in Condition B (with type II errors) than in Condition A (with type I errors). Thus, a self-interested expected payoff maximizer has a relative preference for type II error over type I error. The intuition is that resources are wasted on punishing the innocent in Condition A but not in Condition B. If an innocent person goes to jail, human and social capital is wasted by pointlessly spending resources on imprisonment. In contrast, no such direct costs accrue if a guilty person escapes punishment (e.g. Rizzolli and Saraceno 2013).[4] As is clear from (5), a self-interested income maximizer has a stronger preference for a regime with type II error over one with type I error when the expected costs of sanctions are high, i.e. when sanctions (*s*) are severe and likely (*p*) to be meted out, and when much is at stake (*W*).

[4] There may be indirect costs of type II errors if the wrongfully acquitted commit crimes that they could not have committed if they were in jail. On the other hand, there may also be indirect costs of type I errors if imprisonment makes people more likely to commit crimes (after being released) than they otherwise would have been.

Our design varies two key dimensions of the sanctioning system, namely the severity of sanctions and the probability of error. In particular, free riders pay a fine equal to either 60 or 80 percent of what they put in their private account (i.e. s = 0.6 or s = 0.8). With the high fine, sanctions are deterrent. With the low fine, they are non-deterrent. The probability of error is either 25 or 50 percent. Even when sanctions are theoretically deterrent in the½ absence of errors, they are always theoretically non-deterrent when errors of these magnitudes are introduced. The design therefore allows us to compare errors that render a deterrent sanction non-deterrent to errors that merely weaken an already non-deterrent sanction system.

Predictions for the consequences of judicial error

In the simplest case, i.e. assuming actors are rational and self-interested, standard economic theory predicts full contributions to the public good ($c_i = W$) when sanctions are deterrent and zero contributions otherwise. This implies that judicial error is predicted to have a very strong effect on contributions when errors change sanctions from deterrent to non-deterrent (i.e. with s = 0.8), but to have no effect when the sanction is already non-deterrent in the absence of errors (s = 0.6). The probability of errors should matter for contributions only insofar as it affects whether sanctions are deterrent. Sanctions are non-deterrent at both levels of error-probability we implement (p = 0.25 and p = 0.5) and the effect of judicial errors on contributions should therefore not depend on p.

Behavioral economic theories modify these knife-edge predictions. For example, inequality aversion (as in Fehr and Schmidt 1999) implies that some people derive disutility from free riding when others contribute. Therefore, equilibria with positive contributions may exist, even with non-deterrent sanctions. Such equilibria are more likely to prevail when the marginal per capita return (MPCR) to investment in the public good is high than when it is low. Since the "effective MPCR" ($m + (1 - p)s - 1$ in equations 3 and 4 above) is increasing in s and decreasing in p, the effect of judicial error on contributions is predicted to depend on both sanction severity and error probability. Theories of "reciprocity" (e.g. Rabin 1993) are also consistent with positive contributions in public goods games.

Explaining a pronounced dislike of type I judicial error

The starting point for generating predictions about preferences (more specifically: relative dislikes) for type I vs. type II errors is the notion that type I errors are inherently worse than type II errors. Sayings such as "better that 10 guilty men escape than that one innocent suffer" are attributed to a number of prominent sources (e.g. to William Blackstone, Benjamin Franklin and U.S. Supreme Court Justice Cardozo, see

Grechenig et al. 2010).[5] A preference for type II over type I errors is embedded in all legal systems where the burden of falls on the prosecution rather than the defense.

Table 2 presents results from an incentivized elicitation of norms on the fairness of type I and type II errors among our subjects. To elicit such norms, we follow a procedure developed by Krupka and Weber (2013) and ask a series of questions at the end of the experiment. These questions were framed in the context of the public goods experiment that subjects had just experienced. Subjects were presented with examples of type I and type II errors and asked to state how fair or unfair each type of error was. But the trick of the procedure is not to ask how fair they think the errors are but how fair they think *the average subject* perceives the errors to be. To incentivize such answers, subjects were rewarded (with about 1 USD) if they chose the modal answer in their experimental session. The results clearly confirm our expectations that type I errors are generally viewed as more unfair than type II errors. In particular, 58 percent state that type I errors are either "unfair" or "very unfair", while the corresponding figure for type II errors is only 45 percent. This difference is significant at the 1 percent level (Wilcoxon signed-rank test).

Table 2: Perceived fairness of type I and type II errors (*N* = 190)

	Type I error	*Type II error*
Very fair	4.1	5.5
Fair	4.6	10.5
Somewhat fair	9.1	14.2
Neither fair nor unfair	12.8	9.1
Somewhat unfair	11.4	15.5
Unfair	26.0	22.8
Very unfair	32.0	22.4

Notes: Table presents responses from a questionnaire administered at the end of half of the sessions. The question for type I [type II] error was: "Consider a person who must choose how much of an endowment of 20 points to allocate to a private and to a public account, the same way that you have done in this experiment. Assume that the general rule is that persons who allocate 20 points [zero points] to the public account must pay no fine [16 points in fine] to the experimenter. How fair is it if there is an error, and as a result, one person who allocated 20 points [zero points] to the public account must pay a fine of 16 points [no fine] to the experimenter?"

In principle, a preference for type II over type I errors can be derived both when assuming standard "economic man" preferences and when assuming various kinds of social preferences. Our incentive-compatible mechanism serves to distinguish between these two sources of dislike for type I errors because we measure whether the aggregate willingness to pay exceeds the amount predicted by standard (self-interest) preferences. First, assume standard economic preferences. As noted in eq. (5), expected earnings

[5] But there is clearly no consensus on the 10:1 ratio in the literature, see Volokh (1997).

are *psW* higher with type II errors than with type I errors, all else equal (we discuss the effects of risk and loss aversion below). This fact justifies the hypotheses that a) there is a positive willingness to pay for having type II instead of type I errors, b) this WTP is increasing in *s* and *p*. Second, allow for social preferences. Cox, Servatka and Vadovic (2012) argue that reciprocal reactions to acts of commission are stronger than to acts of omission. The reason is that acts of commission signal a stronger degree of *intent* (to either harm or help) than do acts of omission. A type I error is an act of commission, while a type II error is an act of omission. If people experience more intense utility loss from having their payoff reduced by an act of commission than from an act of omission, then this entails a preference for type II over type I errors.[6]

Rizzolli and Stanca (2012) argue that type I errors have a stronger undermining effect on the so-called "expressive function" of the law than type II errors. The expressive function of the law is to corroborate norms of right and wrong (e.g. Cooter 1998). In our experiment, sanctions corroborate the norm that contributing to the public good is morally right. Type I errors undermine this norm by actively violating it, while type II errors do not directly contradict the norm (because there is no positive reward for free riders, only the absence of punishment). Type I errors may therefore impact more negatively on the deterrence effect of sanctions than type II errors, which may in turn generate a preference for type II over type I errors. If these behavioral mechanisms are important, we expect people's willingness to pay for having type II instead of type I errors to be *higher* than *psW*.

Below, we argue that risk aversion and loss aversion can shape predictions for how much type I errors are disliked in addition to the dislike for type II error. But these effects are not entirely straightforward to predict. As shown below, risk aversion impacts differently depending on the social preference types (e.g. free rider vs. conditional cooperator) and the effect of loss aversion subtly depends on the specification of the reference point.

Since the possibility of judicial error implies uncertainty, risk preferences are potentially important, especially in interaction with social preferences. The effects of errors on contributions depend on the cooperative disposition of the individual (e.g. Fischbacher and Gächter 2010, Thöni et al. 2012). In particular, free riders are not affected by type I errors, because this type of error affects only the "innocent" (free riders pay a fine equal to *sW* both when there is a type I error and when there is no error). So, even when there is a risk of type I errors, the earnings of a free rider are fixed (given her own and other group members' contribution decisions). On the other hand, free riders may gain from type II errors (they pay a fine of *sW* when there is no error and zero when there is). Therefore, when there is a risk of type II

[6] In Cox et al. (2012) experimental subjects' actions result in acts of omission and commission while in ours they originate from the computer (as programmed by the experimenter). It is an open question whether people perceive acts of omission and commission by the computer in the same way that they perceive errors committed by individuals.

errors, the free rider faces a lottery (earnings are higher if the error happens than if it does not). Consequently, the gain from replacing type I errors with type II errors is smaller for a risk-averse free rider than for a risk neutral one. The opposite is true for cooperators. A full cooperator (that is, full contributor to the public good) faces a certain level of income even when type II errors are possible (she pays a fine of zero in both cases) but uncertain income when type I errors may happen (she pays *sW* when an error happens and zero when it does not). Therefore, a risk-averse cooperator has a stronger preference for errors of type II rather than type I than does a risk neutral one. In other words, if agents are risk averse, preferences for type I versus type II errors may depend on cooperativeness: Cooperators have a stronger preference for preventing type I errors than do free riders. For risk neutral individuals, on the other hand, cooperativeness is irrelevant since the expected loss from type I errors relative of type II errors is *psW* regardless of own contribution.

Risk preferences may also have an indirect effect on preferences for different types of errors if they change the way errors affect contributions. As discussed above, expected earnings are always lower (ceteris paribus) with type I than with type II errors. Consider a given decrease in individuals *i*'s contribution to the public good. Assume that parameters are such that sanctions are non-deterrent. The decrease in contributions then leads to an increase in expected earnings. For a risk-averse individual, the increase in utility from this increase in earnings is higher under type I- than under type II errors, simply because the individual operates at a steeper section of the utility function with type I- than with type II errors (see the appendix A for a formal proof of this proposition and Nicita and Rizzolli 2014 for a closely related discussion). For egoistic individuals, the higher utility cost of contributions under type I errors should not matter for the contribution decision because the predicted contribution is zero in any case. However, if agents have social preferences and therefore gain utility from contributing, risk preferences may be important. In particular, if the utility cost of contributing is higher under type I than under type II errors, a stronger dose of social preferences may be needed to induce positive contributions. In that sense, type I errors may have a stronger, negative effect on contributions than type II errors. This should strengthen preferences for type II over type I errors because of the positive externalities of contributions to the public good.[7]

If a utility function displays a kink at the reference point (i.e. the function is strictly steeper immediately below than immediately above the reference point), we say that it displays "loss aversion"

[7] Some subjects may value cooperation not only because it increases earnings but also in its own right. For example, reciprocal subjects may experience strictly higher utility from mutual cooperation than from mutual non-cooperation, even if material payoffs from cooperation and non-cooperation are identical (Rabin 1993). Such preferences should therefore further strengthen preferences for type II over type I errors, to the extent that type II errors make cooperation more likely.

(Kahnemann and Tversky 1979). Rizzolli and Stanca (2012) argue that loss aversion may further increase the negative effects of type I errors, relative to type II errors. But the implications of loss aversion depend on the definition of the reference point. In our experiment, one option is the endowment, *W*. In the standard public goods game, one can ensure a payoff of at least *W* by allocating nothing to public goods production. However, in a system with sanctions, there is no way an individual can be sure to earn at least *W*, even when errors are absent. Another option is to view zero as the reference point. In this case, it is clear that loss aversion implies strengthened preferences for type II over type I errors. The lowest possible earnings in Condition B (type II errors) is min($W(1 - s)$, mW), which (for $s < 1$) is positive. In Condition A, on the other hand, the lowest possible payoff is $W(m - s)$, which is negative for the parameters implemented in this experiment. Hence, period-wise losses are possible in Condition A but not in Condition B. For loss-averse individuals, this strengthens the preference for type II over type I errors.

The upshot of the discussion above is that subjects may prefer type II over type I errors for various reasons. An important aim of the paper is to investigate whether such preferences are in line with risk- and loss neutral income maximization, or whether they are stronger than that, as predicted by theories of risk- and social preferences.

3. Experimental design

The experiment has four phases. Each phase builds on the previous and adds complexity. Phase 1 starts with the basic dilemma situation without sanctions or errors, phase 2 adds (error-free) sanctions, phase 3 adds (exogenous) errors, and in phase 4 we let groups choose between error types. The design thus familiarizes subjects with the basic cooperation problem and the (possibly deterrent) effect of sanctions before they proceed to the more complex settings with errors which are the focus of our attention. At the beginning of each phase, a new set of instructions is handed out and subjects answer control questions on the screen. Also, groups are reshuffled at the beginning of each phase (according to a "stranger matching" protocol).

Phase 1 (no sanctions) is a linear public goods game. Subjects are divided into groups of *n* members. In each period, each subject receives an endowment equal to *W* and decides how much of it to allocate to the production of a public good and how much to keep for themselves. Subject *i*'s payoff function is given by equation (1). Throughout the experiment, we set $n = 5$, $W = 20$ and $m = 0.3$. The payoff function is therefore

$$\pi_i = 20 - c_i + 0.3\sum_{j=1}^{5} c_j \qquad (1')$$

There are five periods in phase 1. In each period, groups are reshuffled.

Phase 2 (no judicial error) is the same as phase 1, except that a) there is only one period, b) subjects now pay a fine of *s* points for each point they allocate to their private account. The payoff function is given by equation (2). With the parameter values implemented, this becomes:

$$\pi_i = (20 - c_i)(1 - s) + 0.3\sum_{j=1}^{5} c_j \qquad (2')$$

s is either 0.8 or 0.6. With *s* = 0.8, the sanction is deterrent. With *s* = 0.6, it is non-deterrent.

Phase 3 (judicial error is possible) is the same as phase 2, except that errors are now introduced. There are two "Conditions" (see Table 1). In Condition A, type I errors are possible, in Condition B, type II errors are possible. If a type I error occurs, a fine of *sW* must be paid regardless of how points are allocated. If type II error occurs, no fine is paid, regardless of how points are allocated. The probability of error, *p*, is either 0.25 or 0.5. Subjects make two decisions in phase 3, i.e. they decide on contributions in both Condition A and Condition B. The choices are made in random order and without feedback.[8]

The payoff functions (in expected terms) in Conditions A and B are, respectively:

Condition A:

$$\pi_i^A = 20(1 - s) + \left(0.3 + (1 - p)s - 1\right)c_i + 0.3\sum_{j \neq i} c_j \qquad (3')$$

Condition B:

$$\pi_i^B = 20\left(1 - (1 - p)s\right) + \left(0.3 + (1 - p)s - 1\right)c_i + 0.3\sum_{j \neq i} c_j \qquad (4')$$

Note that sanctions are non-deterrent (i.e. $0.3 + (1 - p)s - 1 < 0$) for all combinations of *s* and *p* implemented in phase 3 (*s* = 0.6 or *s* = 0.8, *p* = 0.25 or *p* = 0.5). Therefore, in the case of *s* = 0.8, errors change sanctions from deterrent in Phase 2 to non-deterrent in Phase 3. For *s* = 0.6, on the other hand, errors merely weaken an already non-deterrent sanction scheme.

In **Phase 4 (choose type of judicial error)**, subjects participate in deciding whether their group will be in Condition A (type I errors) or in Condition B (type II errors) using an incentive-compatible mechanism (see Messer et al. 2010 and Sausgruber and Tyran 2011). The idea of the mechanism is to elicit the maximum willingness to pay to prevent one type of error while incurring the other type. Thus, the mechanism elicits a "net" (or differential) willingness to pay to be in one condition rather than in the other.

[8] Errors are randomized at individual level, not at group- or session level (i.e. one member of a group may experience an error, while another member of the same group does not). Subjects only receive feedback on their own, personal exposure to error, not the exposure of other group members.

In essence, the mechanism proceeds as follows: groups are randomly assigned to a default condition A or B. The mechanism then elicits the group's maximum willingness to pay (WTP) to prevent being in the default condition and to get the alternative condition instead. Subjects can choose WTP numbers from an interval [-20, 20]. A positive WTP indicates that the group prefers the alternative condition over the default, a negative WTP indicates the converse. Negative WTP mean that subjects demand compensation to get the alternative and such values can be thought of as a willingness to accept (WTA) the alternative condition.

The detailed decision procedure is as follows:

- The group is randomly assigned to a default condition (A or B). Subjects are informed about the default.
- Each group member indicates his/her maximum willingness to pay (WTP) to be in the alternative condition rather than in the default. If the group has been assigned to default Condition B, each subject indicates W_i(A), i.e. the max. WTP to be in Condition A instead of Condition B. In the converse case, subjects indicate W_i(B), i.e. the WTP to be in B rather than A. $W_i(\cdot)$ is a number between -20 and 20, positive numbers indicate that the alternative is preferred.
- The group median $W_{med}(\cdot)$ of the chosen numbers is determined. $W_{med}(\cdot)$ represents the median voter's choice of the WTP to be in the alternative condition rather than in the default condition.
- The computer randomly chooses a price x between -20 and 20 to be in the alternative Condition.
- If $W_{med}(\cdot) > x$, the group gets the alternative condition and pays x. If $x < 0$, each group member receives $-x$ points.
- If $W_{med}(\cdot) \leq x$ the group gets the default Condition and pays/receives nothing.

Participants received feedback at the end of each phase (and at the end of each period in Phase 1). The benefit of this design choice is that feedback is an important means of facilitating full comprehension of experimental rules and incentives among subjects.

After phase 4, we elicit a measure of loss aversion by offering subjects a lottery[9], measure "cognitive reflection" (Fredericks 2005) and, in the second half of experimental sessions, we elicited social norms on aversion to judicial errors in an incentivized task informed by Krupka and Weber (2013), see table 2.

Treatments and participants

The experiment was conducted at the Laboratory for Experimental Economics, University of Copenhagen. A total of 390 freshman economics students, about two months into the program, have participated in the

[9] The choice is between a lottery: "win 15 Danish kroner with probability 0.5, lose 10 Danish kroner with probability 0.5" and a safe option "get nothing". We refer to those rejecting the lottery (19.6 percent) as "loss averse" (see Fehr and Goette 2007).

experiments. About 33 percent of subjects were female. The experiment was implemented with the software z-Tree (Fischbacher 2007).

The eight treatments of our design result from the 2 x 2 x 2 crossing of three dichotomous variables, namely the strength of the sanction *s* (0.6 or 0.8), the probability of error *p* (0.25 or 0.5) and the default in phase 4, i.e. whether subjects stated WTP for being in Condition A rather than B or the converse.

Table 3 shows the number of participants per session along the two main dimensions (table C1 in the appendix details the number of subjects by default in phase 4). We label a treatment with a D (i.e. DH, DL) if the sanction is theoretically deterrent, and with an N (NH, NL) if it is non-deterrent. The letter H stands for high probability of error, and L for low probability of error.

Table 3: Treatments

		Probability of judicial error is	
		High (*p* = 0.5)	**Low** (*p* = 0.25)
Sanction is	**Deterrent** (*s* = 0.8)	**DH** (95)	**DL** (90)
	Non-deterrent (*s* = 0.6)	**NH** (100)	**NL** (105)

Note: The parenthesis below the treatment label shows the number of subjects. The distribution across treatments is uneven because of variation in show-up rates. *N* = 390 in total.

4. Results

Section 4.1 analyzes whether and to what extent errors undermine the deterrent effect of sanctions. Section 4.2 presents results on preferences for type I versus type II errors. Section 4.3 investigates whether the endogenous choice of sanction regime has an effect on contributions to the public good.

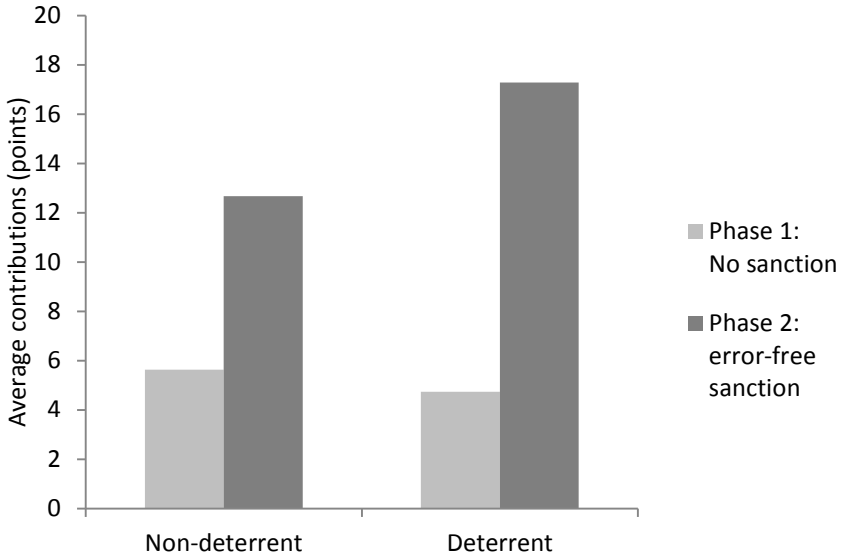
4.1. Do judicial errors undermine cooperation?

As explained in section 3, the design is cumulative and consecutive phases are of increasing complexity. As our main focus is on the consequences of errors and on eliciting relative aversion against errors, we keep the discussion of phases 1 and 2 short. Our main findings from phases 1 and 2 can be succinctly summarized in

Result 1: *In an environment in which a cooperation problem prevails, error-free deterrent sanctions increase efficiency, non-deterrent sanctions do not.*

Phase 1 is a standard linear public goods game without sanctions and our results are in line with what has been found in previous experiments: contributions start at around 40 percent of the endowment and gradually decline with repetition resulting in low contribution rates (see Appendix C for results by period and treatment. As expected, treatment effects are small and mostly insignificant).

Figure 1: Error-free sanctions increase cooperation



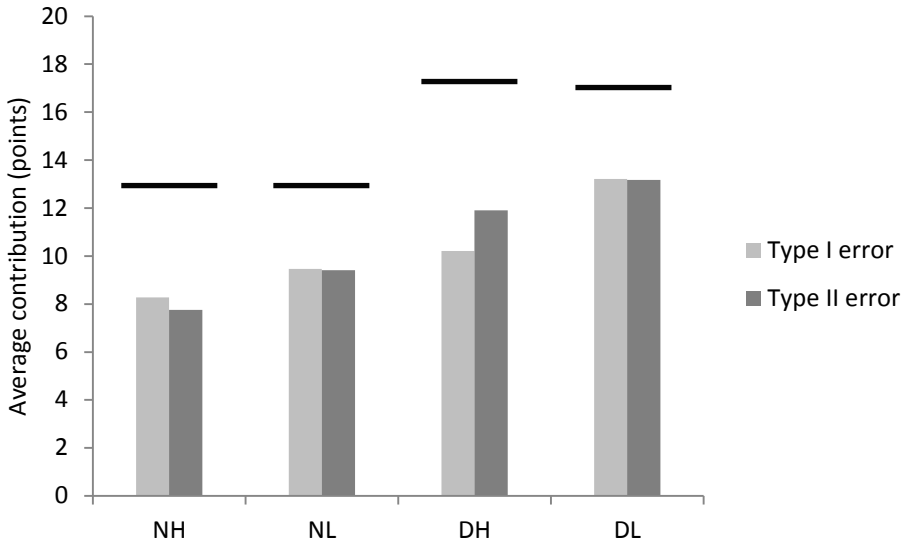
Notes: Light-shaded bars show average contribution in points over all 5 periods of Phase 1. Dark-shaded bars show average contributions in a one-shot contribution choice in which the respective sanction type is administered without error.
N_1 = 205 subjects in Non-deterrent, N_2 = 185 subjects in Deterrent.

Figure 1 shows that imposing error-free sanctions increases contributions. Two facts stand out from this figure. First, deterrent sanctions induce close to full cooperation (the average contribution 17.3). Perhaps more surprising is the fact that theoretically non-deterrent sanctions also improve cooperation, and significantly so. Second, deterrent sanctions are more effective than non-deterrent ones, and only deterrent sanctions increase efficiency. Non-deterrent sanctions induce cooperation rates to about double but deterrent sanctions induce an almost fourfold increase (factor 3.7, from 24% to 86%). Cooperation is significantly higher when sanctions are deterrent than when they are not ($p < 0.01$, MW test). Efficiency

(i.e. taking the cost of sanctions into account) is significantly higher with deterrent sanctions than without sanctions, but efficiency is significantly lower with non-deterrent sanctions than without sanctions.[10]

Figure 2 shows that judicial error undermines cooperation. The black lines in the figure show the cooperation levels when sanctions are error-free, the bars show cooperation levels when errors are present. The resulting loss amounts to between 23 and 42 percent of contributions with error-free sanctions (in phase 2).

Figure 2: Judicial errors undermine cooperation



Notes: Black lines show average contributions absent errors (in phase 2). Bars show average contributions when errors may occur. NH stands for Non-deterrent sanction with High error probability (NL: non-deterrent low), D for deterrent sanctions (see Table 1).

This decline in cooperation is highly statistically significant for both type I and type II errors in all treatments ($p < .01$ in Wilcoxon signed-rank tests for all treatments). While a (strong) effect of errors is predicted by standard theory with deterrent sanctions, the decline with non-deterrent sanctions is perhaps surprising (as is the high level of cooperation in the first place).[11] We summarize these findings in

[10] Non-deterrent sanctions may be counterproductive even if they increase contributions because punishment consumes resources. Comparing phase 2 with the first period of phase 1, earnings are significantly higher when sanctions are deterrent ($p < .01$, WSR test), but significantly *lower* when sanctions are non-deterrent ($p < .01$), a finding that accords well with Tyran and Feld (2006) who also find that imposing non-deterrent sanctions does not increase efficiency.

[11] A possible concern is that the difference in contributions of phase 2 and 3 does not cleanly measure the causal effect of judicial error as we do not provide a difference-in-difference measure (by repeating phase 2 conditions in a control treatment). We speculate that contributions are unlikely to strongly decline in a repetition of phase 2.

Result 2: *Judicial errors undermine the deterrent effect of sanctions. This is the case when errors render a theoretically deterrent sanction scheme non-deterrent. Surprisingly, it is also the case when sanctions merely weaken a theoretically non-deterrent scheme.*

The finding that judicial error undermines the behavioral effects of non-deterrent sanctions contrasts with the standard theory prediction that errors only have an effect if they turn a deterrent sanction scheme into a non-deterrent one. The finding is consistent, however, with studies showing that contributions systematically vary with the marginal per capita return (MPCR) when the MPCR remains strictly below 1 (e.g. Isaac and Walker 1988, see also Ledyard 1995, section 3.3). A higher error probability is equivalent to lowering the effective MPCR. As discussed above, these results may be explained by the influence of social preferences, such as inequality aversion or reciprocity.

Figure 2 also shows that contribution levels are remarkably similar with type I and type II errors, i.e. that the effects of the two types of errors are symmetric – as predicted by standard theory (compare light- and dark-shaded bars). The only exception is DH, where average contributions are 1.7 points higher with type II than type I error ($p < 0.05$ WSR test). In all other treatments and over all conditions jointly, the differences across types of errors are small and statistically insignificant.

Result 3: *The adverse effects of type I and type II errors are symmetric.*

Result 3 is a remarkable finding: Even if people tend to view type I errors as more unfair than type II errors (as we have seen in the survey results shown in table 1 and as we will see below in behavioral data), the two kinds of error affect behavior in the same way. While this finding is in line with standard economic theory, it seems to contrast with results in Dickson et al. (2009), and Rizzolli and Stanca (2012) who report stronger effects of type I than of type II. However, the differences found in these papers are of relatively low magnitude. As explained above, the difference in Dickson et al. (2009) is driven by differences in the behavior of "monitors" (subjects who administered punishment), and not by different reactions to type I and type II errors.

Figure 2 also suggests that cooperation responds to the size of error as expected, i.e. cooperation is undermined more strongly when judicial errors are larger. This impression is confirmed in MW-tests for the difference in effect of errors between treatments with high and low error probabilities. Such tests yield *p*-values of .00 for type I errors and .10 for type II errors.[12]

[12] A manipulation check shows that, as expected, there are no phase 2-differences between those that were assigned to H and L conditions, respectively, in phase 3.

Table 4: Determinants of the effect of errors

	Dep. Var. is change in contributions from phase 2 to 3			
	Type I errors (Condition A)		**Type II errors** (Condition B)	
	(1)	(2)	(3)	(4)
Judicial error probable (p = 0.5)	-2.383***	-2.479***	-1.813*	-2.071**
	(0.907)	(0.955)	(0.923)	(0.982)
Deterrent (s = 0.8)	-1.849**	-1.776**	-0.544	-0.275
	(0.915)	(0.885)	(0.930)	(0.936)
CRT		-0.712*		-0.359
		(0.400)		(0.347)
Female		0.380		1.718
		(1.395)		(1.086)
Loss averse		-0.930		-1.779
		(1.306)		(1.318)
Order (type 2 errors first)		-0.078		0.787
		(1.296)		(1.142)
Contribution in period 1		0.025		0.143*
		(0.081)		(0.081)
Constant	-3.185***	-2.217	-3.845***	-5.873***
	(0.515)	(2.050)	(0.816)	(1.859)
Observations	390	390	390	390
Log likelihood	-1298.0	-1296.40	-1294.30	-1288.9

Notes: Tobit regressions, with censoring from above (20) and below (-20). Standard errors, clustered by session in parentheses. Negative coefficients indicate that cooperation falls from phase 2 to 3 with the variable in question. "Judicial error probable" is a dummy taking the value 1 if p = 0.5. "Deterrent" is a dummy variable taking the value 1 if s = 0.8. "CRT" is Cognitive Reflection Test score. "Loss averse" is a dummy for *not* choosing the lottery described above. "Order" is a dummy for choosing contribution in Condition B before choosing contribution in Condition A in Phase 3. "Contribution in period 1" is contribution in the first period of Phase 1. * significant at 10%; ** significant at 5%; *** significant at 1%.

Table 4 shows that the adverse effect of type I error systematically varies with the deterrence level s and the error probability p, and that the effect of type II error also varies with p but does not vary significantly with s, controlling for other factors. The table shows regressions for the change in contributions from phase 2 to 3, i.e. it statistically explains differences in the effect of introducing errors separately for type I and II errors (recall that subjects made choices for both types of error in phase 3 in random order and without feedback). We estimate tobit regressions to take censoring into account (however, OLS regressions generate qualitatively similar results).

Regression models (1) and (2) estimate the effect of type I errors (Condition A) while models (3) and (4) concern type II errors (Condition B). Models 1 and 3 include only the treatment variables (dummies

taking the value 1 for p = 0.5 and for s = 0.8, respectively), and models 2 and 4 add controls for cognitive reflection (CRT), gender, loss aversion, individual contribution to the public good in period 1, and an indicator for the sequence in which subjects made their choices.

Results show a robust effect of error probability in all models, i.e. that cooperation is undermined more strongly when judicial error is high than when it is low (see first line in table 4). There is a significant effect of the severity of the sanction ("deterrence") in Condition A, i.e. type I errors undermine cooperation more when sanctions are deterrent than when they are not. But this is not significantly the case with type II errors. The controls do not change the estimates of the treatment effects much, and they do not explain the adverse effect of judicial errors either. We summarize these results in

Result 4: *Judicial errors are more harmful when the error probability is high. Errors tend to undermine cooperation more when sanctions are deterrent than when they are not, but the evidence for this effect is strong only for type I errors.*

Is it better to have error-prone sanctions than no sanctions at all? The answer to that question is likely to depend on parameter values but for the values implemented here (with rather high probabilities of error), and taking Period 1 of Phase 1 as our benchmark, we find that the answer is negative, at least when judged by subjects' average earnings. We find that total payoff is *lower* with error-prone sanctions (in phase 3) than without sanctions (i.e. in period 1 of phase 1). This difference is highly significant in all treatments and conditions, with the only exception of DH in Condition B, according to Wilcoxon signed-rank tests. Hence, a (seriously) flawed sanction scheme is not necessarily better than no sanctions at all, and might even be counterproductive.

4.2. Results on preferences for type I versus type II errors

We now turn to results on endogenous choice of error type. This section discusses the relative dislike for the two types of error. We discuss the consequences of having chosen one or the other type of error in the next section.

We measure the relative dislike for the two types of errors by eliciting a "net" (or differential) willingness to pay (WTP) to be in a regime with type II error rather than in one with type I error, and vice versa (see section 3). The prediction is that rational, self-interested and risk-neutral agents are willing to pay $WTP_I = psW$ to prevent the regime with type I error (and get a regime with type II error instead) because subjects must endure pointless (i.e. wasted) costly punishment with type I error (the subscript indicates the default). Half of the subjects are assigned to the treatment in which the default is type I error, the other half are assigned to a treatment in which the default is type II error. In this case, subjects effectively indicate a willingness to accept (WTA) being in a regime with type I error instead. It is a WTA

because a rational, risk-neutral agent prefers type II error, and therefore he will demand compensation to be in a regime with type I error. In fact, the minimum amount he has to be paid to agree to being in a regime with type I error when the default is type II error is $WTA_{II} = -WTP_{II} = psW$. That is, the prediction is that $WTP_{I} = - WTP_{II} = WTA_{II} = psW$.[13]

The reason we elicit two "net" (or differential) WTP measures are as follows. An alternative to our procedure would have been to assign subjects to a default without errors and ask them to state their WTA for errors of type I and type II, respectively. In addition, we could have assigned others to a regime with (type I or type II) errors as a default and asked them to state the WTP to be in regime without error. Our procedure has the advantage that only two (rather than four) treatments are necessary. Perhaps more importantly, our procedure avoids the endowment effect which tends to create a discrepancy between WTA and WTP measures by always assigning a "bad" as a default (see Brenner et al. 2007 who show that endowment effects can be reversed with bads). Our procedure also helps us to detect and control for other biases. For example, if some participants have a tendency to name positive rather than negative numbers, a discrepancy WTP > WTA would be observed.

***Figure 3*:** WTP for type II error and WTA for type I error

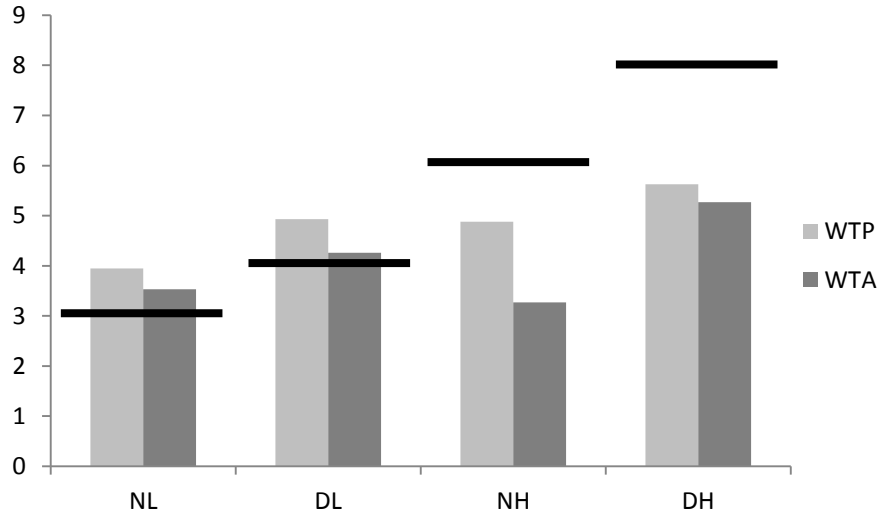


Figure 3 shows mean WTP (bars with light shading) and WTA (dark shading) by treatment along with the predicted values for how much a regime with type II errors is preferred over type I errors (black lines). Four facts stand out. First, the measures are all highly significantly bounded away from zero. This means

[13] Note that even though we distinguish between WTP for type II errors and WTA for type I errors here, wording in experimental instructions is entirely symmetric across treatments with, respectively, type I and type II errors as the default; see appendix B. In practice, therefore, WTA was measured as a negative willingness to pay.

that our experiment provides clear and strong support for the notion that subjects dislike type I error more than type II error (or, that they prefer type II error over type I error).[14] The finding is robust as it is obtained across different levels of deterrent vs. non-deterrent and moderately vs. highly error-prone sanctions. Second, this systematic (relative) dislike of type I error is robust to the default used in eliciting preferences, i.e. whether WTA or WTP is elicited. We find that the elicited measures are very similar across defaults, i.e. whether we measure WTA or WTP yields similar results. There is no significant difference in any single treatments or overall (MW tests).

Result 5: *There is strong evidence that subjects dislike type I errors more than type II errors, all else equal.*

The third finding that stands out from figure 3 is that the measures are very well in line with predictions (with slight tendency for a preference towards type II errors) when the error probability is low, i.e. in NL and DL. None of these deviations from the predictions are statistically significant (*t*-tests). Fourth, for the treatments with high error probability, the measures are clearly below predictions. We think the correct interpretation is that subjects underreact to (very) high error probabilities. In fact, while WTA and WTP measures are predicted to double when the error probability doubles, they only increase very moderately (on average by 14.3%). Support for this interpretation comes from Table 5.[15]

Table 5: Relative preference of a regime with type II over type I errors

	(1) WTP*= *psW*	(2) All	(3) *t*-test	(4) CRT $\geq$ 2	(5) *t*-test
DH	8	5.5	0.010***	6.7	0.441
DL	4	4.6	0.502	5.0	0.351
NH	6	4.2	0.045**	4.5	0.202
NL	3	3.8	0.380	1.8	0.381

Notes: *N* = 390. Columns 2 and 4 show average WTP/WTA values. Columns (3) and (5) show p-values for t-tests. The corresponding null hypothesis is H0: WTP = WTP*.

Table 5 shows, in line with Figure 3, that there is a clear preference for type II over type I error, and that the extent of this preference is in line with predictions from standard theory, in particular for cognitively sophisticated subjects. Column 1 shows WTP*, the predictions for the strength of the preference of type II over type I errors according to standard theory. Column 2 shows the average values (we merge the WTA and WTP values here) in our treatments when considering all subjects. Column 3

[14] Recall that subjects choose WTP/WTA values between -20 and 20, i.e. the choice set was balanced between positive and negative values.

[15] Table C2 in the appendix presents regression analyses of the determinants of WTP/WTA.

shows that these averages are not significantly different from predictions for the low error probabilities, but are significantly below prediction for the high error probabilities (DH and DL) according to *t*-tests.

Table 5 also presents a test of whether confusion might explain this unexpected deviation from standard predictions. The experimental environment is somewhat complicated, and despite our design that facilitates understanding by gradually building complexity, and even though we were extremely careful about administering detailed instructions and comprehension tests, some confusion may have persisted. Conceivably, confused subjects bias their answers toward zero, which is in the middle of the support (-20, 20) and which might be perceived as a "neutral" choice. Such confusion would then falsely be interpreted as a preference for type I over type II error. One way to deal with this issue is to exploit the data on cognitive reflection (we administered a test for cognitive reflection at the end of the session). The last two columns of Table 5 show results for the 175 subjects (45 percent) who scored 2 or 3 on the CRT test (scores range from 0 to 3, see Fredericks 2005). In particular, column 5 shows that elicited preferences of the cognitively more sophisticated subjects are fully in line with the predictions of standard theory (no significant difference to WTP* according to *t*-tests). We summarize these findings in

Result 6: *The hypothesis that income maximization explains the strength of preferences for type II over type I errors cannot be rejected. We do not find evidence in support of an additional, psychological bias against type I errors.*

This finding is surprising because it runs counter to the widely-held intuition that there is a dislike of type I errors above and beyond what is prescribed by simple income maximization.

4.3. Effects of endogeneity

A growing body of experimental research shows that institutions to overcome social dilemmas have more beneficial effects when they are chosen through voting than when they are imposed exogenously (Tyran and Feld 2006, Dal Bo, Foster and Putterman 2010, Sutter, Haigner and Kocher 2010, Baldassari and Grossman 2011, Markussen, Putterman and Tyran 2014, Kube et al. 2014).[16] Inspired by this literature, we now investigate whether judicial errors have different consequences for cooperation behavior when subjects have been involved in choosing which type of error to prevent than when they have not.

One explanation for why there might be such a "dividend of democracy" is related to signaling. For example, Tyran and Feld (2006) argue that voting for (error-free) non-deterrent sanctions is a signal of cooperativeness, and reciprocal subjects respond to such a belief by cooperating. However, such an account seems less plausible in our setting than in the settings where such a "dividend" has been found because

[16] On the other hand, Markussen, Reuben and Tyran (2014) find no such effect in a public goods experiment where subjects vote about introducing competition between groups.

voting as such is not directly observed in our design, because voting is not a simple and clear dichotomous (yes or no) decision but a continuous choice, and because the outcome is importantly determined by randomness in our design. We investigate potential effects of endogeneity by comparing cooperation in phase 3 and 4 of the experiment. Errors are exogenous in phase 3 but endogenous in phase 4.

Table 6: Change in contributions from phase 3 (type of error is imposed) to phase 4 (type of error is chosen)

	Condition A (**type I error**)				Condition B (**type II error**)		
Treatment	(2) Avg. change	(3) *t*-test	(4) WSR test		(5) Avg. change	(6) *t*-test	(7) WSR test
NH	0.02	0.98	0.75		-0.40	0.68	0.65
NL	1.47	0.08*	0.10		-1.93	0.08*	0.07*
DH	2.13	0.07*	0.10*		-1.02	0.12	0.38
DL	1.82	0.06*	0.02**		0.77	0.56	0.54
All	1.33	0.01***	0.01***		-0.83	0.10*	0.16

Notes: the columns labeled t-test and Wilcoxon signed-rank (WSR) test show *p*-values. In both cases, the Null hypothesis is that the difference is equal to 0.

Table 6 shows that the effects of endogeneity depend on the type of error: choice of type I errors increased contributions from phase 3 to 4, but choice of type II error tended to reduce contributions. In particular, column 2 shows that contributions increase in all treatments and that contributions increase in Condition A (type I error is imposed in phase 3 and type I error is chosen in phase 4) overall by 1.33 units from phase 3 to 4. Column 3 shows that this overall increase is highly significant (bottom line), and that the increase is significant in three out of four treatments according to *t*-tests. In Condition B (when type II errors are imposed in phase 3 and chosen in phase 4), the picture is reversed. There is now a drop in contributions overall (by -0.83 units) and in three of four treatments. The overall decline is weakly significant in a *t*-test, but not in a WSR test, and it is significant in one treatment according to both tests. As an alternative test, we can simply compare the contributions in phase 4 across conditions A and B. They are about 1.7 points higher when type I is chosen than if type II is chosen (p = 0.03, MW test). The data thus suggests that there is a "dividend of democracy" for one type of error but not the other: endogenous implementation of type I error seems to increase contributions, while endogenous implementation of type II errors seems to decrease them.[17] But it is of course not clear that the effect is causal, it might well be a selection effect.

[17] Results for earnings go in the same direction as those for contributions and are statistically significant when all treatments are pooled (p = 0.09 for type I errors, p = 0.04 for type II errors, WSR tests). The treatment-specific differences are significant in one case for type I errors (DL) and in three cases for type II errors (NL, DH, DL), although one of these (DL) goes in the opposite direction (positive) than in the other treatments and overall.

Table 7: Determinants of contributions to public good in phase 4

	Dependent variable: Contribution to PG in phase 4		
Deterrent (s = 0.8)	9.370***	10.133***	10.219***
	(1.672)	(1.480)	(1.506)
Error prob. (p = 0.5)	-3.021*	-5.028***	-5.043***
	(1.598)	(1.630)	(1.613)
Default for WTP is type I error	-0.629	-0.410	-0.405
	(1.963)	(1.655)	(1.641)
Condition B (type II errors)	-3.415	-3.583*	-4.209*
	(2.279)	(2.139)	(2.225)
WTP for type II errors	0.036	0.008	0.006
	(0.115)	(0.112)	(0.112)
Experienced Type 1 error in phase 3		1.629	1.554
		(1.363)	(1.328)
Experienced Type 2 error in phase 3		2.710*	2.689*
		(1.571)	(1.583)
CRT		-1.109*	-1.171**
		(0.577)	(0.544)
Female		0.937	0.824
		(1.845)	(1.836)
Loss averse		-1.882	-3.624
		(1.810)	(2.418)
Contribution in period 1		0.567***	0.568***
		(0.094)	(0.094)
Loss averse*contribution in per. 1			3.126
			(2.979)
Constant	9.513***	5.781***	6.206***
	(1.560)	(1.796)	(1.746)
Observations	390	390	390
Log likelihood	-914.9	-900.5	-900.2

Notes: Tobit regressions, allowing for censoring above (20) and below (0). Standard errors, clustered by session, in parentheses. "Judicial error probable" is a dummy taking the value 1 if p = 0.5. "Deterrent" is a dummy variable taking the value 1 if s = 0.8. "Default for WTP is type I error" is a dummy for choosing WTP_I rather than WTA_{II}. "Condition B (type II errors)" is a dummy for being in Condition B in Phase 4. "WTP for type II errors" is WTP_I when type I errors is the default and WTA_{II} when type II errors is the default. "Experienced type I (II) errors in Phase 3" are dummies for having actually been exposed to type I (II) error in Phase 3. "CRT" is Cognitive Reflection Test score. "Loss averse" is a dummy for *not* choosing the lottery described above. * significant at 10%; ** significant at 5%; *** significant at 1%

One possibility for the causal interpretation is this: choosing type II errors means to express support for the possibility that there is no punishment of free riders. Expressing a preference for type II errors (high

WTP) may therefore be interpreted as a signal that a subject intends to free ride. This interpretation might induce other group members to lower contributions if they have reciprocal preferences. Conversely, expressing support for type I error (low WTA) implies an acceptance of punishment, which may signal a willingness to contribute to the public good. An alternative explanation of the observed "dividend of democracy" is that it is driven by selection. Suppose low contributors have high WTP to be in regime with type II error. Groups that happen to have many low contributors would thus be more likely to select into condition B.

Table 7 estimates determinants of contributions to the public good in phase 4, and serves to shed light on the selection hypothesis. Explanatory variables include both an indicator for being in Condition B and the subject's WTP for being in Condition B. If the effect of endogeneity is driven by selection, we would expect that individuals with high WTP (i.e. high preference for type II errors) contribute less than others, and that being in condition B is insignificant, conditional on WTP. In fact, the opposite is true. Being in Condition B has a negative effect on contributions, significant at the 10-percent level in regressions 2 and 3, even after controlling for WTP, which is insignificant. Selection is therefore unlikely to explain our result, whereas we cannot reject the hypothesis that the beneficial effects of endogeneity are driven by how subjects interpret other subject's WTP choices.

5. Conclusion

This paper has developed a framework to investigate both the aversion against judicial error and the effects of such errors on cooperation. An important innovation of our design is that it enables us to measure not only whether subjects prefer type II errors (letting the guilty go) over type I errors (punishing the innocent), but also how strong such preferences are.

Concerning the effects of judicial errors, our results show that errors are harmful since they significantly undermine the deterrent effect of sanctions. The harm done by type I and type II errors is symmetric, and the harm done by judicial errors is stronger when errors are more frequent. It is also stronger when errors turn a theoretically deterrent sanction scheme into a theoretically non-deterrent one than when they merely weaken an already non-deterrent sanction, but this difference is much smaller than predicted by standard theory and is only statistically significant for type I errors.

Concerning the aversion against judicial error, our results also show that subjects have a clear preference for type II over type I errors. We infer such preferences from voting in an incentive-compatible mechanism. We derive the prediction for how a rational, self-interest, risk-neutral individual votes. Such a

voter has preference for being in a regime with type II error over type I error because type I error involves pointless but costly punishing. While an aversion above and beyond what is implied by a motive to maximize income seems plausible by various accounts, we find no evidence for such an aversion. From this perspective, the problem with type I errors is simply that they generate a loss of resources for pointless punishment.

Are these results inconsistent with findings such as those reported in the introduction, which show a stronger social norm of aversion to type I- than to type II errors? Not necessarily. Such social norms may simply reflect internalization of the resource losses generated by type I errors. Alternatively, latent feelings related to type I error aversion may not be evoked by the experimental setting implemented here. Future research should investigate, for example, whether framing experimental instructions more explicitly in terms of "guilt" and "innocence," or having penalties which are of more than monetary consequence, generate stronger observable aversion to type I errors than seen here.[18]

In line with a recent stream of experimental literature we find evidence in support of a "dividend of democracy". Type I errors are less harmful, and type II errors are more harmful when subjects participate in deciding which type of error to prevent than when regimes with such errors are exogenously imposed. The reason may be that expressing a high preference for type II errors is interpreted by other subjects as a sign of intentions to free ride, because type II errors offer free riders a chance of escaping without sanctions.

In general, some aspects of the results presented are well in line with standard, economic theory, while others point to the importance of social preferences. Both the strength of preferences for type II errors over type I errors and the symmetric effects of both error types on contributions to the public good are remarkably well predicted by standard theory. On the other hand, the deterrence-undermining effect of errors depends much less than predicted by standard theory on whether errors change sanctions from deterrent to non-deterrent. Also, the extent to which errors undermine the deterrent effect of sanctions significantly depends on error probability, which was not predicted by standard theory, given our choice of parameter values. In addition, positive contributions to the public good were observed in many cases when sanctions were non-deterrent. Thus, a full understanding of how formal sanctions, including ones subject to errors, affect the inclination to contribute to a public good, would draw not only on the analytics of rational self-interest but also on behavioral approaches, including ones that allow for the presence of other-regarding preferences.

[18] Perhaps the strongest aversion to type I errors stems from the thought that an innocent person could spend years in prison or even be put to death, consequences for which no monetary compensation seems adequate. Comparable non-monetary penalties would be difficult to simulate in an experiment, although penalties having the form of irrecoverably lost opportunities could conceivably be engineered.

References

Abbink, K. (2006): Laboratory experiments on corruption. In Rose-Ackerman, S. (ed.): *International Handbook on the Economics of Corruption*. Cheltenham: Edward Elgar: Ch. 14.

Alm, J., McClelland, G.H. and Schulze, W.D. (1992): Why Do People Pay Taxes? *Journal of Public Economics* 48: 21-38.

Ambrus, A. and Greiner, B. (2012): Imperfect Monitoring with Costly Punishment: An Experimental Study. *American Economic Review* 102(7): 3317-3332.

Andreoni, J. and Gee, L.K. (2012): Gun for Hire: Delegated Enforcement and Peer Punishment in Public Goods Provision. *Journal of Public Economics* 96(11): 1036-1046.

Baldassari, D. and Grossman, G. (2011): Centralized Sanctioning and Legitimate Authority Promote Cooperation in Humans. *Proceedings of the National Academy of Sciences* 108: 11023-11027.

Brenner, L., Rottenstreich, Y., Sood, S. and Bilgin, B. (2007): On the Psychology of Loss Aversion: Possession, Valence, and Reversals of the Endowment Effect. *Journal of Consumer Research* 34(3): 369-376.

Carpenter, J. (2007): Punishing Free-Riders: How Group Size Affects Mutual Monitoring and the Provision of Public Goods. *Games and Economic Behavior* 60(1): 31-52.

Carpenter, J., Kariv, S. and Schotter, A. (2012): Network Architecture, Cooperation and Punishment in Public Goods Experiments. *Review of Economic Design* 16: 93-118.

Chaudhuri, A. (2010): Sustaining Cooperation in Laboratory Public Goods Experiments: A Selective Survey of the Literature. *Experimental Economics* 14(1): 47-83.

Cinyabugama, M., Page, T. and Putterman, L. (2006): Can Second-Order Punishment Deter Perverse Punishment? *Experimental Economics* 9(3): 265-79.

Cooter, R. (1998): Expressive Law and Economics. *Journal of Legal Studies* 27: 585–608.

Cox, J.C., Servatka, M. and Vadovic, R. (2012): Status Quo Effects in Fairness Games: Reciprocal Responses to Acts of Commission vs. Acts of Omission. Working paper, Georgia State University.

Dai, Z., Hogarth, R.M. and Villeval, M.C. (2014): Ambiguity on Audits and Cooperation in a Public Goods Game. IZA Discussion Paper No. 7932.

Dal Bó, P., Foster, A. and Putterman, L. (2010): Institutions and Behavior: Experimental Evidence on the Effects of Democracy. *American Economic Review* 100(5): 2205-2229.

DeAngelo, G. and Charness, G. (2012): Deterrence, expected cost, uncertainty and voting: Experimental evidence. *Journal of Risk and Uncertainty* 44: 73-100.

Dickson, E.S., Gordon, S.C. and Huber, G.A. (2009): Enforcement and Compliance in an Uncertain World: An Experimental Investigation. *Journal of Politics* 71: 1357-1378.

Ertan, A., Page, T. and Putterman, L. (2009): Who to Punish? Individual Decisions and Majority Rule in Mitigating the Free-Rider Problem. *European Economic Review* 53(5): 495-511.

Fehr, E. and Gächter, S. (2000): Cooperation and Punishment in Public Goods Experiments. *American Economic Review* 90(4): 980-994.

Fehr, E. and Goette, L. (2007): Do Workers Work More if Wages Are High? Evidence from a Randomized Field Experiment. *American Economic Review* 97(1): 298-317.

Fehr, E. and Schmidt, K.M. (1999): A Theory of Fairness, Competition, and Cooperation. *Quarterly Journal of Economics* 114(3): 817-868.

Fischbacher, U. (2007): z-Tree: Zurich toolbox for ready-made economic experiments. *Experimental Economics* 10: 171-178.

Fischbacher, U. and Gächter, S. (2010): Social Preferences, Beliefs, and the Dynamics of Free Riding in Public Goods Games. *American Economic Review* 100(1): 541-556.

Fredericks, S. (2005): Cognitive Reflection and Decision Making. *Journal of Economic Perspectives* 19(4): 25-42.

Garoupa, N. (1997): The Theory of Optimal Law Enforcement. *Journal of Economic Surveys* 11(3): 267-295.

Grechenig, K., Nicklisch, A., and Thöni, C. (2010): Punishment despite Reasonable Doubt – A Public Goods Experiment with Sanctions under Uncertainty. *Journal of Empirical Legal Studies* 7(4): 847-867.

Grechenig, K., Nicklisch, A., and Thöni, C. (2013): Information-sensitive Leviathans: The Emergence of Centralized Punishment. Unpublished manuscript.

Gürerk, Ö., Irlenbusch, B. and Rockenbach, B. (2009): Motivating Teammates: The Leader's Choice Between Positive and Negative Incentives. *Journal of Economic Psychology* 30: 591-607.

Harbaugh, W.T., Mocan, N.H. and Visser, M.S. (2011): Theft and Deterrence. NBER working paper 17059.

Harris, J.R. (1970): On the Economics of Law and Order. *Journal Political Economy* 18(1): 165-174.

Heijden, E.v.d., Potters, J., and Sefton, M. (2009): Hierarchy and Opportunism in Teams. *Journal of Economic Behavior and Organization* 69: 39-50.

Herrmann, B., Thöni, C. and Gächter, S. (2008): Antisocial Punishment Across Societies. *Science* 319(5868): 1362-1367.

Kahneman, D. and Tversky, A. (1979): Prospect Theory: An Analysis of Decision under Risk. *Econometrica* 47(2): 263-292.

Kamei, K., Putterman, L. and Tyran, J.-R. (forthcoming): State or Nature? Formal vs. Informal Sanctioning in the Voluntary Provision of Public Goods. *Experimental Economics (in press)*.

Khadjavi, M. (2014): On the Interaction of Deterrence and Emotions. Journal of Law, Economics and Organization. Forthcoming.

Krupka, E.L. and Weber, R.A. (2013): Identifying Social Norms using Coordination Games: Why does Dictator Game Sharing Vary? *Journal of the European Economic Association* 11(3): 495-524.

Kube, S., Schaube, S., Schildberg-Hörisch, H. and Khachatryan, E. (2014): Institution Formation and Cooperation with Heterogeneous Agents. IZA DP No. 8533

Ledyard, J.O. (1995): Public Goods: Some Experimental Results. In J. Kagel and A. Roth (eds.): *Handbook of Experimental Economics*. Princeton: Princeton University Press: Ch. 2.

Markussen, T., Putterman, L., and Tyran, J.-R. (2014): Self-Organization for Collective Action: An Experimental Study of Voting on Sanction Regimes. *Review of Economic Studies* 81(1): 301-324.

Markussen, T., Reuben, E. and Tyran, J.-R. (2014): Competition, Cooperation, and Collective Choice. *Economic Journal* 124(574): 163-195.

Messer, K.D., Poe, G.L., Rondeau, D., Schulze, W.D. and Vossler, C.A. (2010): Social Preferences and Voting: An Exploration using a Novel Preference Revealing Mechanism. *Journal of Public Economics* 94: 308-317.

Nicita, A. and Rizzolli, M. (2014): In Dubio Pro Reo. Behavioral Explanations of Pro-defendant Bias in Procedures. CESifo Economic Studies.

Nosenzo, D. and Sefton, M. (2014): Promoting Cooperation: The Distribution of Reward and Punishment Power. In: Van Lange, P.A.M., Rockenbach, B. and Yamagishi, T. (eds.): *Reward and Punishment in Social Dilemmas*, Oxford: Oxford University Press.

O'Gorman, R., Henrich, J. and Van Vugt, M. (2009): Constraining Free Riding in Public Goods Games: Designated Solitary Punishers Can Sustain Cooperation. *Proceedings of the Royal Society B*, 276: 323-329.

Png, I.P.L. (1986): Optimal Subsidies and Damages in the Presence of Judicial Error. *International Review of Law and Economics* 6: 101-105.

Polinsky, M. and Shavell, S. (2000): The Economic Theory of Public Enforcement of Law. *Journal of Economic Literature* 38: 45-76.

Putterman, L., Tyran, J.-R. and Kamei, K.(2011): Public Goods and Voting on Formal Sanction Schemes. *Journal of Public Economics* 96(9-10): 1213-1222.

Rabin, M. (1993): Incorporating Fairness into Game Theory and Economics. *American Economic Review* 83(5): 1281-1302.

Rabin, M. (2000): Risk Aversion and Expected Utility Theory: A Calibration Theorem. *Econometrica* 68(5): 1281-1292.

Rizzolli, M. and Stanca, L. (2012): Judicial Errors and Crime Deterrence: Theory and Experimental Evidence. *Journal of Law and Economics* 55: 311-338.

Rizzolli, M. and Saraceno, M. (2013): Better That Ten Guilty Persons Escape: Punishment Costs Explain the Standard of Evidence. *Public Choice* 155: 395-411.

Sausgruber, R. and Tyran, J.-R. (2011): Are We Taxing Ourselves? How Deliberation and Experience Shape Voting on Taxes. *Journal of Public Economics* 95(1-2): 164-176.

Schildberg-Hörisch, H. and Strassmair, C. (2012): An Experimental Test of the Deterrence Hypothesis. *Journal of Law, Economics, and Organization* 28(3): 447-459.

Sonnemans, J. and van Dijk, F. 2011. Errors in Judicial Decisions. Experimental Results. *Journal of Law, Economics and Organization* 28(4): 687-716.

Sutter, M., Haigner, S. and Kocher, M. (2010): Choosing the Stick or the Carrot? – Endogenous Institutional Choice in Social Dilemma Situations. *Review of Economic Studies* 77(4): 1540-1566.

Tan, F. and Yim, A. (2014): Can Strategic Uncertainty Help Deter Tax Evasion? An Experiment on Auditing Rules. Journal of Economic Psychology 40: 161-174.

Thöni, C., Tyran, J.-R. and Wengström, E. (2012): Microfoundations of Social Capital. *Journal of Public Economics* 96(7): 635-643.

Traulsen A., Röhl, T. and Milinski, M. (2012): An Economic Experiment Reveals that Humans prefer Pool Punishment to maintain the Commons. *Proceedings of the Royal Society B* 279(1743): 3716-3721.

Tyran, J.-R. and Feld, L.P. (2006): Achieving Compliance when Legal Sanctions are Non-deterrent. *Scandinavian Journal of Economics* 108(1): 1-22.

Yamagishi, T. (1986): The Provision of a Sanctioning System as a Public Good. *Journal of Personality and Social Psychology* 51(1): 110-116.

Volokh, A. (1997): n guility men. *Pennsylvania Law Review* 146(1): 173-216.

Zhang, B., Li, C., De Silva, H., Bednarik, P., and Sigmund, K. (2014): The Evolution of Sanctioning Institutions: An Experimental Approach to the Social Contract. *Experimental Economics* 51(2): 285-303.

Appendix A

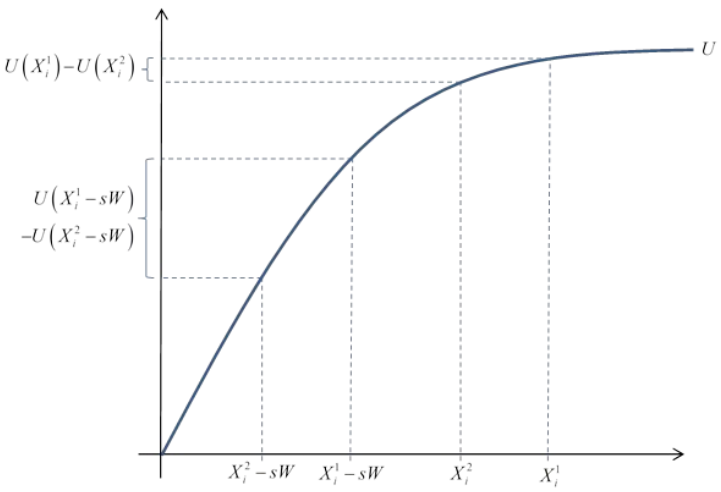
Appendix A provides a proof that for a risk averse individual, the utility gain from decreasing his/her contribution to PG is higher under type I (Condition A) than under type II errors (Condition B).

Assume that agent *i* has the concave utility function $U(\pi_i)$ and maximizes expected utility.

Consider a decrease in contributions from c_i^2 to c_i^1, $c_i^1 < c_i^2$. Denote the expected utility gain in Condition A by ΔU_i^A and in Condition B by ΔU_i^B. Define $X_i^1 = W - c_i^1 + mc_i^1 + m\sum_{j\neq i} c_j$ and $X_i^2 = W - c_i^2 + mc_i^2 + m\sum_{j\neq i} c_j$. Note that $X_i^1 > X_i^2$. We have that:

$$
\begin{aligned}
\Delta U_i^A - \Delta U_i^B = {} & pU\left(W(1-s) - c_i^1 + mc_i^1 + m\sum_{j\neq i} c_j\right) + (1-p)U\left(\left(W - c_i^1\right)(1-s) + mc_i^1 + m\sum_{j\neq i} c_j\right) \\
& - \left(pU\left(W(1-s) - c_i^2 + mc_i^2 + m\sum_{j\neq i} c_j\right) + (1-p)U\left(\left(W - c_i^2\right)(1-s) + mc_i^2 + m\sum_{j\neq i} c_j\right)\right) \\
& - \left(pU\left(W - c_i^1 + mc_i^1 + m\sum_{j\neq i} c_j\right) + (1-p)U\left(\left(W - c_i^1\right)(1-s) + mc_i^1 + m\sum_{j\neq i} c_j\right)\right) \\
& + \left(pU\left(W - c_i^2 + mc_i^2 + m\sum_{j\neq i} c_j\right) + (1-p)U\left(\left(W - c_i^2\right)(1-s) + mc_i^2 + m\sum_{j\neq i} c_j\right)\right) \\
= {} & pU\left(W - c_i^1 + mc_i^1 + m\sum_{j\neq i} c_j - sW\right) - pU\left(W - c_i^2 + mc_i^2 + m\sum_{j\neq i} c_j - sW\right) \\
& - \left(pU\left(W - c_i^1 + mc_i^1 + m\sum_{j\neq i} c_j\right) - pU\left(W - c_i^2 + mc_i^2 + m\sum_{j\neq i} c_j\right)\right) \\
= {} & p\left(U\left(X_i^1 - sW\right) - U\left(X_i^2 - sW\right)\right) - p\left(U\left(X_i^1\right) - U\left(X_i^2\right)\right) > 0
\end{aligned}
$$

The inequality holds by concavity of *U* (see illustration below).



Appendix B: Experimental instructions, treatment DL2

Welcome. You are now taking part in an economics experiment. Depending on your decisions and the decisions of other participants, you will be able to earn money. How you can earn money is described in these instructions. Please read them carefully. You will have to answer control questions to check that you understand the instructions. You can only continue the experiment when you have answered these questions correctly.

During the experiment you are not allowed to communicate with other participants. If you have a question, please raise your hand. One of us will come to answer your question. Sometimes you may have to wait a short while before the experiment continues. Please be patient.

During the experiment your earnings will be calculated in points. In the first phase of the experiment, points will be converted to Danish kroner at the following rate:

8 points = 1 DKK

At the end of the experiment your total earnings will be paid out to you in cash.

The experiment has four phases. The following instructions explain the details of phase 1. The details of the subsequent phases will be explained later.

Instructions for Phase 1

Phase 1 has **5 periods**. At the beginning of each period, all participants are randomly divided into **groups of 5**. This means that you are in a group with four other participants and that in any given period you will almost certainly not be matched with the same participants as in the previous period. Nobody knows which other participants are in their group, and nobody will be informed who was in which group after the experiment.

In each period, each group member, yourself included, will be given **20 points**. In each period you will have to make one decision.

Your decision

You and the four others in your group simultaneously decide how to use the 20 points. There are two possibilities:

1. **You can allocate points to a group account.**
2. **You can allocate points to a private account.**

You will be asked to indicate the number of points you want to allocate to the group account. Only integers between 0 and 20 are allowed for this purpose. The remaining points will automatically be allocated to your private account. Your earnings depend on the total number of points in the group account, and the number of points in your private account.

How to calculate your earnings

Your earnings from your private account are equal to the number of points you allocate to it. That is, **for each point you allocate to your private account you get 1 point as earnings**. For example, your earnings from the private account equal 3 points if you allocate 3 points to it. The points you allocate to your private account do not affect the earnings of the others in your group.

Your earnings from the group account equal the **sum** of points allocated to the group account by all 5 group members multiplied by 0.3. **For each point you allocate to the group account you and all others in your group each get 0.3 points as earnings**. For example, if the sum of points in the group account is 30, then your earnings from the group account and the earnings of each of the others in your group from the group account are equal to 9 points.

Your earnings can be calculated with the following formula:

20 – (points you allocated to the group account) + 0.3 * (sum of points allocated by all group members to the group account)

Note that you get 1 point as earnings for each point you allocate to your private account. If you instead allocate 1 extra point to the group account, your earnings from the group account increase by 0.3 * 1 = 0.3 points and your earnings from your private account decrease by 1 point. However, by allocating 1 extra point to the group account, the earnings of the other 4 group members also increase by 0.3 points. Therefore, the total group earnings increase by 0.3 * 5 = 1.5 points. Note that you also obtain earnings from points allocated to the group account by others. You obtain 0.3 * 1 = 0.3 points for each point allocated to the group account by another member.

Example

Suppose you allocate 10 points to the group account, the second and third members of your group each allocate 20 points to the group account, and the remaining two individuals allocate 0 points each. In this case, the sum of points in the group account is 10 + 20 + 20 + 0 + 0 = 50 points. Each group member gets earnings of 0.3 * 50 = 15 points from the group account.

Your total earnings are: 20 – 10 + (0.3 * 50) = 10 + 15 = 25 points.

The second and third members' earnings are: 20 – 20 + (0.3 * 50) = 0 + 15 = 15 points.

The fourth and fifth members' earnings are: 20 – 0 + (0.3 * 50) = 20 + 15 = 35 points.

Do you have any questions? (Please raise your hand.)

Instructions for Phase 2

Please read these instructions carefully. Again, you will have to answer control questions to check that you understand the instructions.

In the rest of the experiment, points will be converted to Danish kroner at the following rate:

1 point = 1 DKK

In phase 2, everything is the same as in the previous phase of the experiment, with one exception: Each individual now pays a fine equal to **80 percent** of the amount of points allocated to the **private account.**

Groups of five are randomly formed at the beginning of this phase. That is, you are now in a new group. Phase 2 has only one period.

Earnings are now calculated as follows:

20 – (points you allocate to group account) + 0.3*(sum of points allocated by all in group to group account) – 0.8*(points you allocate to private account)

For example, suppose you allocate 10 points to the group account, the second and third members of your group each allocate 20 points to the group account, and the remaining two individuals allocate 0 points

each. In this case, the sum of points in the group account is 10 + 20 + 20 + 0 + 0 = 50 points. Each group member gets earnings of 0.3 * 50 = 15 points from the group account.

Your total earnings are: 20 - 10 + (0.3 * 50) – (0.8*10) = 10 + 15 – 8 = 17 points.

The last term (0.8*10) indicates the fine you pay for allocating 10 points to the private account. Notice that for each point you put in your private account, you gain 0.2 points (that is, you gain 1 point as income and lose 0.8 points in fines); and for each point you put in the group account, you gain 0.3 points.

Instructions for Phase 3

Please read these instructions carefully. Again, you will have to answer control questions to check that you understand the instructions.

Phase 3 is like the previous one in that groups of five individuals are randomly formed. That is, you are now in a new group. Again, you make decisions about allocating 20 points to either a private account or a group account.

Everything is the same as in phase 2, with the exception that fines can be subject to **error**. In the previous phase, there were no errors in the sense that an allocation to the private account was ALWAYS fined and an allocation to the group account was NEVER fined. Now, two types of errors may occur: In Type 1 error, BOTH the private and the public account are subject to fines. In Type 2 error, NEITHER allocations to the private nor the group account are fined. We explain the details of what the errors mean next.

Type 1 error: BOTH accounts are fined. When Type 1 error occurs, BOTH the points you allocate to the **private account** and the points you allocate to the **group account** are fined. The fine per point is the same as in the previous phase. Thus, when this type of error occurs, you pay 80 percent on all the 20 points you are given, which equals 16 points. Note that in case of Type 1 error, the fine you pay is a fixed amount (16 points) and the fine you pay is **independent** of how you allocate your points to the two accounts. Thus, when Type 1 error occurs, you are fined for allocating points to the group account.

When a Type 1 error occurs, earnings are calculated as follows:

20 – (points you allocate to group account) + 0.3*(sum of points allocated by all in group to group account) – 0.8*(points you allocate to private account) - 0.8*(points you allocate to group account)

= 20 – (points you allocate to group account) + 0.3*(sum of points allocated by all in group to group account) - 16

Notice that when a Type 1 error occurs, for each point you put in your private account, you gain 0.2 points; and for each point you put in the group account, you *lose* 0.5 points (that is, you gain 0.3 points as income and pay 0.8 points in fines).

Type 2 error: NEITHER account is fined. When Type 2 error occurs, you pay NO fines for either your allocations to the private or the group account. Note that in case of Type 2 error, the fine you pay is zero points **independent** of how you allocate your points to the two accounts. Thus, when Type 2 error occurs, you are not fined for allocating points to the private account.

When a Type 2 error occurs, earnings are calculated in the same way as in Phase 1 of the experiment:

20 – (points you allocated to the group account) + 0.3 * (sum of points allocated by all group members to the group account)

Notice that when a Type 2 error occurs, for each point you put in your private account, you gain 1 point; and for each point you put in the group account, you gain 0.3 points.

Your task

You are now asked to make allocation choices in two conditions, called **Condition A** and **Condition B**. The conditions differ by the **chance** that each type of error occurs. (In the absence of an error, the conditions are identical as in the previous phase.)

Condition A: 25% chance of Type 1 error. A **Type 2 error never occurs**. This means that in 25% of the cases, BOTH accounts are fined and you have to pay 16 points in fines independent of your allocation choice. In 75% of the cases only the points you put to the private account are fined and the points you put into the group account are not fined. So, with a 25% chance you pay a fine of 16 independent of how you allocate your points, and with a 75% chance the situation is the same as in the previous phase.

Condition B: 25% chance of a Type 2 error. A Type 1 error never occurs. This means that in 25% of the cases, NEITHER account is fined and you pay zero fines independent of your allocation choice. In 75% of the cases only the points you put to the private account are fined and the points you put into the group account are not fined. So, with a 25% chance you pay no fine independent of how you allocate your points, and with a 75% chance the situation is the same as in the previous phase.

Table 1 summarizes the different conditions:

Table 1: Conditions

	Type 1 error (fine is 16 independent of allocation choice) occurs with a chance of	**Type 2 error** (fine is 0 independent of allocation choice) occurs with a chance of	**No errors** (fine is 0.8 points per point allocated to the private account; no fine for points allocated to the group account) occur with a chance of
Condition A	25%	0%	75%
Condition B	0%	25%	75%

The table shows that in Condition A there is a 25% chance that everyone pays a fine of 16 points no matter how they choose. In Condition B there is a 25% chance that nobody pays a fine no matter how they chose.

In both conditions, there a 75% chance that no error occurs. That is, there is a 75% chance in both conditions that only allocations to the private account are fined and allocations to the group account are not fined.

How we proceed: Each participant decides how many points to allocate to the group account for both Conditions A and B. The choices will appear in random order. When all participants have made their choices for both conditions, the computer determines randomly according to the odds indicated in the table above whether a particular type of error occurs. Then, the computer calculates the incomes and fines and you are informed about the results for both conditions.

Instructions for Phase 4

Please read these instructions carefully. Again, you will have to answer control questions to check that you understand the instructions.

Again, groups of five are randomly formed at the beginning of this phase. That is, you are now in a new group.

Preview of this phase

In this phase, your group will first decide whether to make the allocation decision under Condition A or B. You will then learn which Condition has been chosen for your group and make the allocation decision between private and group account as before.

The table below recaps what Conditions A and B are. Recall that in Condition A there is a 25% chance that everyone pays a fine of 16 points no matter how they choose. In Condition B there is a 25% chance that nobody pays a fine no matter how they chose.

In both conditions, there a 75% chance that no error occurs. That is, there is a 75% chance in both conditions that only allocations to the private account are fined but allocations to the group account are not fined.

	Type 1 error (fine is 16 independent of allocation choice) occurs with a chance of	**Type 2** error (fine is 0 independent of allocation choice) occurs with a chance of	**No errors** (fine is 0.8 points per point allocated to the private account; no fine for points allocated to the group account) occur with a chance of
Condition A	25%	0%	75%
Condition B	0%	25%	75%

Choosing between Condition A and Condition B

You and the other members of your group now participate in **deciding which condition, A or B**, your group will be in. When the choice is made, you will then make one single allocation decision as before. Here is how the decision process works:

First, each group member states the **maximum** amount of points he or she is willing to pay to be in **Condition B** instead of **Condition A.** Call the amount you choose "x". You can choose any amount between **-20 and 20 points.** Choosing a negative amount means that you prefer Condition A over Condition B. A positive amount means that you prefer Condition B over Condition A. The more you prefer Condition B over Condition A, the higher the amount you should choose. The more you prefer Condition A over Condition B, the larger the *negative* number you should choose.

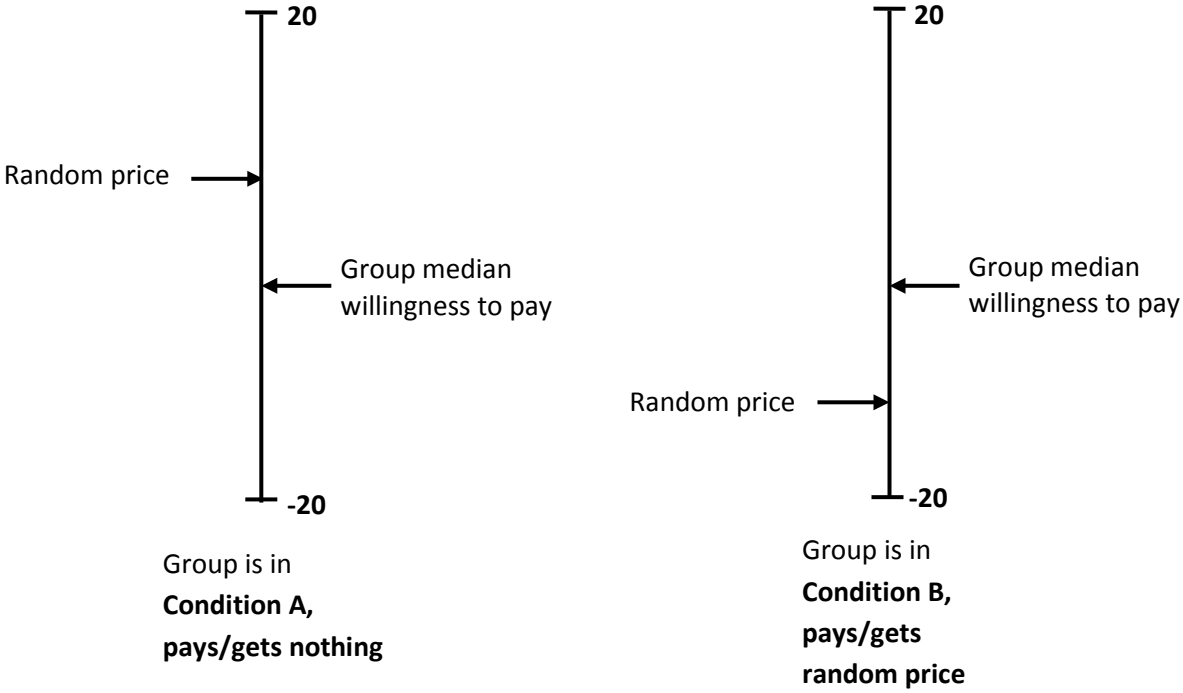
Second, after all group members have chosen x, the x's are ranked from lowest to highest. The third number in this ranking (the "median") is defined as the maximum price your *group* is willing to pay (per member) for being in Condition B instead of Condition A, or the "maximum willingness to pay." For example, suppose the five members of your group choose x's as follows: -10, -2, 3, 9, 18. Then, your group's maximum willingness to pay for being in Condition B instead of Condition A is 3.

Third, the computer randomly determines a price between -20 and 20 for being in Condition B instead of Condition A. All prices in this interval are equally likely. Now, two things can happen.

If the price determined by the computer is *lower* than your group's maximum willingness to pay, your group will use **Condition B**, and each group member pays *the price determined **by the computer*** for using Condition B rather than Condition A. This means that if your group gets Condition B, you do *not* pay your group's maximum willingness to pay to get B but less (except if the computer draws exactly the median x).

If the price determined by the computer is *higher* than your group's maximum willingness to pay, your group will be in **Condition A**. You will pay nothing, since you do not get to be in Condition B instead of Condition A. Figure 1 illustrates how the procedure works.

Figure 1



In the left panel, the price happens to be higher than what the group is willing to pay for getting Condition B rather than Condition A. Thus, the group does not get Condition B. It gets Condition A and does not pay or receive any additional amount.

In the right panel, the price happens to be lower than what the group is willing to pay for getting Condition B rather than Condition A. Thus, the group gets Condition B. Each group member pays the price (if the random price is positive) or receives money (if the random price is negative). Note that the group pays or receives the random price in this case and not the maximum willingness to pay stated.

Here is an **example**: Suppose your group's maximum willingness to pay for being in Condition B instead of Condition A is 3.

a) Suppose the price randomly determined by the computer is **-12**. Since 3 is higher than -12, your group will be in Condition B and each member will *receive* 12 points (the same as paying -12 points) in addition to your other earnings.

b) Suppose the price randomly determined by the computer is **2**. Since 3 is higher than 2, your group will be in Condition B and each member will *pay* 2 points (a deduction from your other earnings).

c) Suppose the price randomly determined by the computer is 10. Since 3 is lower than 10, your group will be in Condition A and you will not pay or receive any points in addition to your other earnings in the period.

Here is another example. Suppose your group's maximum willingness to pay for being in Condition B instead of Condition A is –7.

a) Suppose that the price randomly determined by the computer is –9. Since –9 is smaller than –7, your group will be in Condition B and "pay the price" of –9, which means you will *receive* 9 points in addition to your other earnings.

b) Suppose the price randomly determined by the computer is, say, -5, your group will be in Condition A and you will not pay or receive anything in addition to your other earnings.

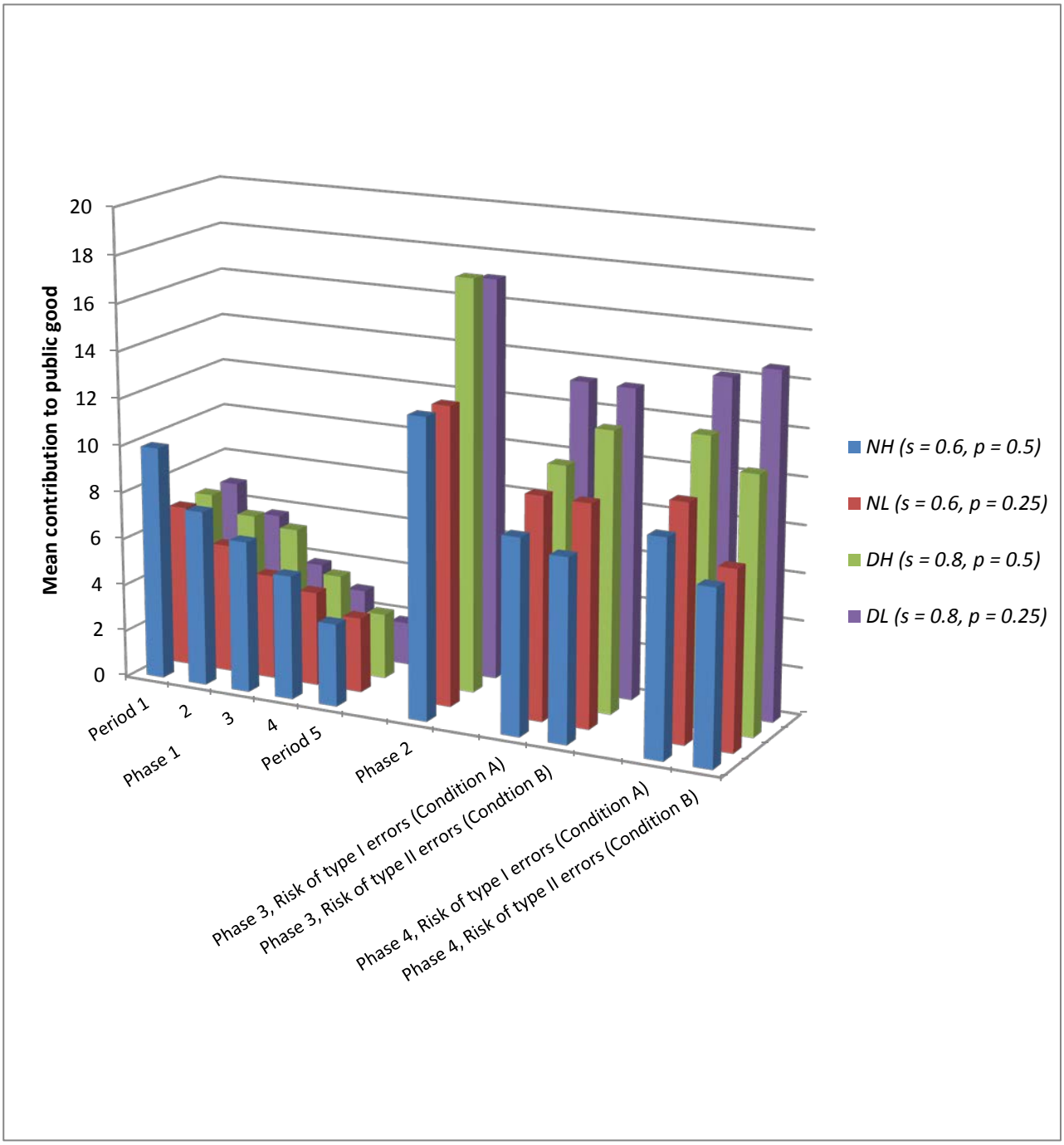
Intuitively speaking, stating a positive number x means that you prefer Condition B over A, stating a negative number x means that you prefer Condition A over B. If you prefer B over A, you are willing to pay at most x to obtain B (that is, to avoid A). If you prefer A over B, you must be compensated to obtain B. In either case, you should choose the amount x such that you are **indifferent** between A and B, given the maximum price you pay or the maximum compensation you receive.

Note that the amount x you state does not directly affect how much you have to pay or receive but it affects **how likely** you are to obtain either A or B. The higher the median x in your group, the more likely it is that your group gets Condition B. The lower the median x is in your group, the more likely it is that your group gets Condition A. The price each group member pays is determined by the random draw. For example, suppose the median x in your group is 20. The group is certain to receive Condition B and each group member pays at most 20 but the average payment is zero (because all prices between -20 and 20 are equally likely). Suppose instead the median x in your group is 0. The group receives Condition B with 50% chance and Condition A with 50% chance. Each group member pays at most 0 but on average receives 10.

When choosing your maximum willingness to pay for being in Condition B instead of Condition A (x), think about how much YOU prefer condition B over A and state that value. The value you state should not be affected by what you think others may choose or what you expect the random price to be. You are always best off if you simply state the amount YOU prefer. To give an example, suppose you are willing to pay up to 2 points to be in Condition B, but state that you are only willing to pay up to -1 points. Suppose your stated value of -1 happens to be the group median and that the computer draws a random price equal to 0 points. Then your group will be in Condition A, even though you could have been in your preferred condition B for a price of 0 points, which was lower than your true willingness to pay (2). You would have been better off stating your true value of 2 rather than stating a lower value (here: -1). The same logic applies to stating higher values than your true x. In any case, you are always best off to truthfully state the value you prefer.

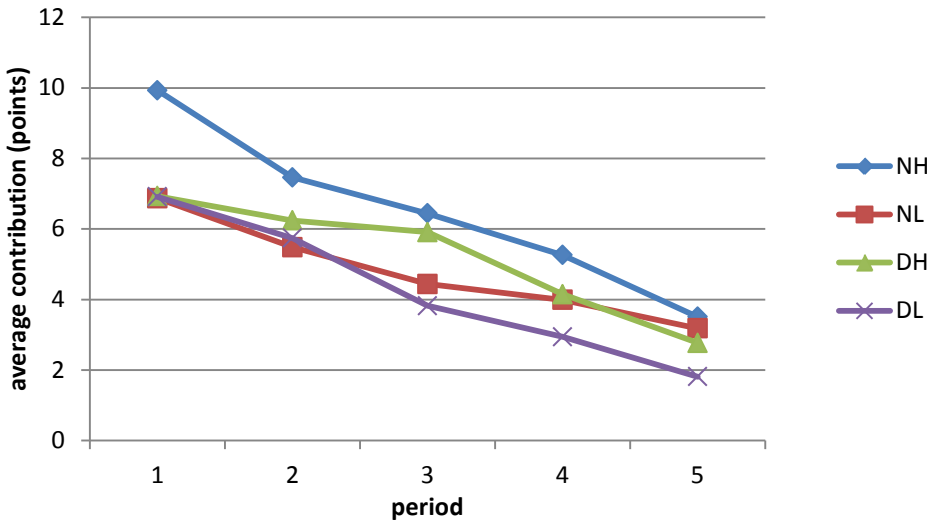
Appendix C: Additional tables and figures

Figure C1: Overview of contributions



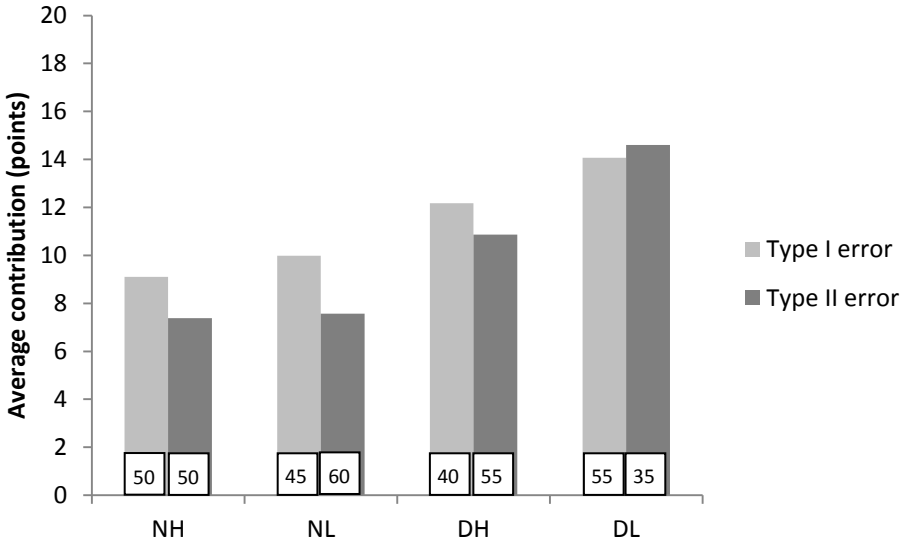
Notes: Number of observations by treatment in phases 1 to 3 is given by Table 3 in the main text. The respective numbers for phase 4 are in Condition A/Condition B: NH: 50/50, NL: 45/60, DH: 40/55, DL: 55/35.

Figure C2: Average contribution in phase 1



Notes: See Table 3 for treatment definitions and number of observations. Pairwise treatment comparisons of average contributions in Phase 1 yield the following p-values (MW tests): DH vs. DL: .12, NL vs. NH: .03, NH vs. DH: .12, NL vs. DL: .25, DH vs. NL: .58, NH vs. DL: .00. Note that significant differences are driven by high contributions in NH and low contributions in DL. This pattern is exactly the opposite of what we expect, and observe, in Phases 2-4. Therefore, the treatment effects observed in Phases 2-4 are not explained by the (small) differences observed in Phase 1.

Figure C2: Contributions in Phase 4



Note: Numbers of observations are indicated at the bottom of each bar.

Table C1: Treatments and participants

	Error probability = 0.5		Error probability = 0.25	
	Default is Type II	Default is Type I	Default is Type II	Default is Type I
Deterrent sanctions (s = 0.8)	DH1 (40)	DH2 (55)	DL1 (45)	DL2 (45)
Non-deterrent sanctions (s = 0.6)	NH1 (40)	NH2 (60)	NL1 (45)	NL2 (60)

Strength of relative preference of type II over type I error

Here we use regression analyses to investigate whether WTP for type II errors and WTA for type I errors vary with the treatment variables in the way predicted by theory. We pool data on WTP and WTA and refer to both as "WTP". Figure 3 and Table 5 in the main text show that WTP is indeed increasing with the severity of sanctions and the probability of error, as expected. Table C2 presents regression analyses that investigate whether these effects are statistically significant and robust to the inclusion of control variables. Again, we estimate tobit regressions. OLS regressions yield similar results. The dependent variable is WTP. All regressions are run both for the full set of subjects and for those with CRT ≥ 2.

Models 1 and 2 include only the treatment dummies, including an indicator for type I errors being the default condition. Models 3 and 4 control for gender, risk/loss aversion, contribution in period 1 and CRT score. Based on the analysis in section 2, which showed that the effect of risk aversion may depend on subject type (free rider or cooperator), the interaction between risk/loss aversion and contribution in period 1 is also included. Models 5 and 6 add controls for being hit by type I and type II errors, respectively, in phase 3. Since errors are random, experience of error should not affect WTP for a rational agent, but it may still be the case that subjects feel most strongly about a particular type of error if they have actually been exposed to it.

Results show that the effects of sanction severity and error probability are always positive, as predicted. Error probability is only significant among cognitively sophisticated subjects, while sanction severity is significant at the 10-percent level in the full set of subjects and at the five-percent level among the cognitively sophisticated. This is consistent with the view that there is less confusion among these subjects.

Among subjects with CRT ≥ 2, there is also a significant, negative effect of type I errors being the default condition. There are no significant effects of gender, loss aversion, contribution in period 1 or, as

mentioned above, CRT score. Experience of type II error in phase 3 has a positive effect on WTP for type II errors. Note that a positive effect is what we would expect since exposure to type II error is *beneficial* from the exposed individuals point of view, in the sense of avoiding payment of a fine.

Table C2: Determinants of willingness to pay for type II instead of type I errors

	Dependent variable: WTP					
	All	CRT ≥ 2	All	CRT ≥ 2	All	CRT ≥ 2
	(1)	(2)	(3)	(4)	(5)	(6)
Deterrent (*s* = 0.8)	1.184*	2.735**	1.292*	2.696**	1.152	2.599**
	(0.697)	(1.147)	(0.761)	(1.119)	(0.741)	(1.104)
Judicial error probable (*p* = 0.5)	0.781	3.105***	0.786	3.057***	0.003	2.469*
	(0.747)	(1.097)	(0.812)	(1.084)	(0.853)	(1.255)
Default for WTP is type I error	0.674	-4.484***	0.712	-4.429***	0.657	-4.618***
	(0.742)	(1.101)	(0.757)	(1.009)	(0.749)	(0.994)
CRT			-0.072	-0.724	-0.046	-0.471
			(0.765)	(1.829)	(0.773)	(1.825)
Female			0.096	1.463	0.024	1.428
			(1.337)	(1.962)	(1.261)	(1.895)
Loss averse			-1.816	-0.172	-1.615	0.239
			(1.761)	(3.107)	(1.828)	(3.066)
Contribution in period 1			-0.009	0.076	0.003	0.093
			(0.051)	(0.092)	(0.054)	(0.093)
Loss averse * contribution in per. 1			0.041	-0.447	-0.001	-0.489
			(0.219)	(0.359)	(0.230)	(0.354)
Experienced type 1 error in phase 3					0.292	-0.837
					(1.146)	(1.379)
Experienced type 2 error in phase 3					2.596**	2.514*
					(1.095)	(1.331)
Constant	3.524***	4.308***	3.895**	5.579	3.223*	4.608
	(0.760)	(1.130)	(1.748)	(3.859)	(1.945)	(4.035)
Log likelihood	-1328.1	-584.0	-1327.3	-581.8	-1324.6	-580.3
Observations	390	175	390	175	390	175

Notes: Tobit regressions, allowing for censoring above (20) and below (-20). Standard errors, clustered by session, in parentheses. WTP is defined as WTA for those who had type II errors as the default. "Deterrent" is a dummy variable taking the value 1 if *s* = 0.8. "Judicial error probable" is a dummy taking the value 1 if *p* = 0.5. "Default for WTP is type I error" is a dummy for choosing WTP_I rather than WTA_{II}. "CRT" is Cognitive Reflection Test score. "Loss averse" is a dummy for *not* choosing the lottery described above. "Experienced type I (II) errors in Phase 3" are dummies for having actually been exposed to type I (II) error in Phase 3.* significant at 10%; ** significant at 5%; *** significant at 1%.