Let the Punishment Fit the Criminal: An Experimental Study*

Josef Montag† James Tremewan‡

We use a laboratory experiment to study the extent to which people tailor levels of punishment to the subjective experience of the person to receive that punishment, for both monetary and non-monetary sanctions. We find that subjects tend to apply higher fines to wealthier individuals. Additionally, subjects assign more repetitions of a tedious task to those with a lower willingness to pay to avoid it. We find no evidence that the distributions of monetary and non-monetary punishments are different when considered as proportions of the maximum possible punishment, but that this does not hold when non-monetary punishments are converted into monetary equivalents. This suggests that subjects do not have in mind a particular level of disutility from the punishment, but rather are guided by the sentencing possibilities.

JEL classification: C91, D03, K14, K40.
Keywords: monetary punishment, non-monetary punishment, subjective experience, laboratory experiment.
This paper is forthcoming in the *Journal of Economics Behavior & Organization*.

1 Introduction

The maxim that the punishment should fit the crime, first postulated in Cicero's *De Legibus* (*On the Laws*) in 106 BC, has been ever since a core principle of criminal justice. A more

*For helpful comments and criticisms we would like to thank two anonymous reviewers, the editor, Richard Boylan, Cornelius Cappelen, David Gray, Thomas de Haan, Adam Kolber, Alfredo Paloyo, Hans Pitlik, Martin Rode, Roee Sarel, as well as seminar participants at the Czech Academy of Sciences and the University of Vienna. This research was supported by the Heinrich Graf Hardegg'sche Stiftung.

†International School of Economics, Kazakh-British Technical University. Tole bi st. 59, 050000 Almaty, Kazakhstan. Email: josef.montag@gmail.com.

‡Department of Economics, Vienna University of Economics and Business and Department of Economics, University of Auckland. Email: james.tremewan@auckland.ac.nz.

open question, however, is whether punishment should fit the criminal. Keeping a nominal punishment constant, the punishees' experience will inevitably vary: an individual's cost of serving a number of years in a prison, for instance, will depend on their psychological characteristics as well as lost opportunities to enjoy life outside prison, such as wealth and personal relationships. An analogous argument can be made with respect to monetary punishment, which will affect individuals differently as a result of decreasing marginal utility of money, among other factors. As a consequence, different individuals will *experience* punishment differently.[1]

Besides judges and juries, a range of other public actors make punishment decisions relevant to economic outcomes in non-market settings. Boards governing professionals such as doctors, nurses, and lawyers can temporarily or permanently suspend the right to practice of those who violate ethical norms. Athletes and sports people are frequently barred from competing for taking banned substances, missing drug tests, or bringing their sport into disrepute by their on-field or off-field behaviour. Schools and colleges suspend students from attending class for a variety of non-criminal activities. Finally, in countries such as the United Kingdom, case workers can temporarily suspend benefit payments if the beneficiary fails to meet certain conditions. In all these instances, the responsible authorities have latitude for determining the degree of punishment, and may or may not take into account the subjective experience of those they are punishing.

This paper is the first to use an incentivized laboratory experiment in order to ascertain the role of punishees' subjective experience in punishment determination, and in this way it contributes to several literatures. First, we add to the experimental literature on third party punishment. Second, we contribute to the current debate among legal scholars and

[1]See Becker (1968, p. 195), noting that "if the monetary value of the punishment by, say, imprisonment were independent of income, the length of the sentence would be inversely related to income, because the value placed on a given sentence is positively related to income." Similarly, Posner (1985, p. 1212) notes that "[t]he economic objection to punishing by inflicting physical pain is not . . . that people have different thresholds of pain that make it difficult to calibrate the severity of the punishment—imprisonment and death are subject to the same problem." See also Polinsky and Shavell (1984). For recent evidence on individual-level variability of pain perception see Schulz et al. (2012). For studies of subjective experience of imprisonment see DeVeaux (2013); Edney (2004); Raaijmakers et al. (2017); Williams et al. (2013). For evidence on post-release effects of incarceration see Fazel and Baillargeon (2011); Lott (1992); Massoglia (2008); Massoglia et al. (2011); Pager (2003).

philosophers concerned about the role of subjective effects of punishment. Third, we add to the law and economics literature on the use of monetary and non-monetary punishments and deterrence of heterogeneous offenders.

There are two main arguments as to why the subjective experience of punishment should be taken into account. First, in one retributive view of justice, the level of suffering inflicted by a punishment should be in proportion to the gravity of the crime. Second, from a deterrence perspective, the optimal level of a sanction depends on the potential criminal's disutility of that sanction; a constant nominal punishment leads to ineffective underdeterrence, or costly overdeterrence depending on the individual.

On the other hand there are theories of punishment which do not equate punishment with suffering (Gray 2010), notwithstanding the fact that assigning different levels punishments to different people for an identical offence may be viewed as ethically unacceptable, violating the principle of equality before the law (Markel and Flanders 2010). Indeed, the United States Sentencing Guidelines Manual discourages judges from calibrating sentences based on convicts' subjective experience (Kolber 2009b).

Whether or not one holds that subjective experience of punishment should be taken into account, it is important to understand how the general public views punishment and justice. An inconsistency between people's values and preferences on the one hand and the existing policies or proposed reforms on the other may well be a problem. Specifically, if the public's understanding of justice was such that punishment should only be fitted to the crime, those who believe that punishment should be co-determined by the level subjective discomfort felt by the punishee might find their policy recommendations resisted. Similarly, for those to whom justice requires equality of nominal punishment, a finding that individuals do use signals about subjective experience to adjust the magnitude of the punishment they assign would be worrisome in the context of punishment decision-making where the authority is given lee-way in sentencing; a natural tendency to make such adjustments would need to be countered to achieve just outcomes.

An important consideration is the possibility that views on the acceptability of calibrating punishments may depend on the type of sanction: many people find the idea of conditioning fines on a convict's income reasonable, but the idea of similarly tailoring prison sentences distasteful (Montag and Sobek 2014). Indeed, predetermined fines in some jurisdictions are proportional to income (e.g. speeding fines in Finland and Switzerland) and judges have been shown to vary discretionary fines according to a convict's wealth (Donna and Espín-Sánchez 2015). However, we are unaware of any similar circumstances with regard to non-monetary punishments.

In this paper we use a laboratory experiment to investigate whether people find it acceptable to condition punishments on the punishees' subjective experience. It differs from Montag and Sobek's (2014) earlier vignette experiment in two methodologically important aspects: (i) We elicit the subjects' true valuation of punishment and provide this information to individuals who make punishment decisions. In the earlier study, the degree to which the different fictional convicts would suffer differentially from punishment was not made explicit and may not have been clear to subjects. (ii) The punishment decisions made by our subjects have real consequences for the individuals who are to be punished, as is the case with real-world adjudicators. This should increase the likelihood of establishing subjects' true preferences towards punishment as they have to deliberate their punishment decisions more carefully.

In contrast to Montag and Sobek (2014), we find that subjects take into account subjective experience not only for monetary punishments (where the fines they assign are increasing in the punishee's wealth), but also for non-monetary punishments (where the assigned number of repetitions of a tedious task is decreasing in the elicited willingness to pay to avoid the task). We find no difference between monetary and non-monetary treatments in the levels of punishments subjects assign when measured as a proportion of the maximum possible sanction; however, when the non-monetary punishments are measured in terms of their monetary equivalents, the distribution of punishment levels differ significantly, monetary punishment being harsher. This suggests that while subjects

take into account subjective experience of punishees in a relative sense, they do not have in mind an absolute level of disutility they wish to inflict for a given offence. Rather, subjects are guided by the sentencing possibilities. This result highlights the role of statutory limitations and sentencing guidelines in shaping punishment decisions.

2 Related Literature

There is an extensive experimental literature on punishment. One strand of the literature examines the deterrent effect of different exogenous punishment regimes, where the size of fines and probability of detection is exogenously manipulated by the experimenter (e.g. Alm 2012 and Rizzolli and Stanca 2012). A second strand considers endogenous punishment, where the decision to punish, and sometimes the level of punishment, is determined by subjects themselves. Punishment decisions can be made either by "second parties" who are directly affected by the potential punishee's choices, or "third parties" who are not materially impacted. This paper belongs to the endogenous third party punishment strand.

One important question is what motivates third party punishment. Punishment is typically triggered by violation of a norm, for example fairness or cooperation (Fehr and Fischbacher 2004), or truth-telling (Ohtsubo et al. 2010). Carpenter and Matthews (2012) demonstrate a role for indignation in mediating the response of the third party to a norm violation, while Nelissen and Zeelenberg (2009) show that either anger or guilt is sufficient. Charness et al. (2008) and Leibbrandt and López-Pérez (2012) show that third parties punish to equalise payoffs of the first and second parties, and the latter also finds evidence that spite plays a role (for more evidence of purely spiteful punishment see Abbink and Herrmann 2011).

There is robust evidence of "antisocial" punishment whereby subjects are punished for acting pro-socially in a public goods game (e.g. Herrmann et al. 2008), which could result from spiteful/competitive preferences or norm-based reasons. People have also been

found to act spitefully towards those who do better than them in a real effort task (Granic and Wagner 2017).

The degree of punishment may be influenced by the actions of the offender, or characteristics of the punisher or offender. In terms of actions, a common finding is that the punishment is increasing in the magnitude of the violation, for example the size of the deviation from a 50/50 split in a dictator game (Bernhard et al. 2006), or from the average group contribution in a public goods game (Carpenter and Matthews 2009).

Characteristics of the punisher that have been identified as affecting punishment decisions are culture (Henrich et al. 2006) and gender (Alatas et al. 2009; Fišar et al. 2016). The only studies we are aware of that consider how characteristics of the offender relate to levels of punishment look at group membership: punishment is lower when the offender is from the same tribe (Bernhard et al. 2006) or caste (Hoff et al. 2011) as the punisher. Montag and Sobek (2014), already discussed, appears to be the only existing experimental work relating subjective experience of the offender to punishment levels.

Apart from the experimental literature on punishment, this paper aims to contribute to a current debate among legal scholars and philosophers about the subjective aspects of punishment. The debate itself is perhaps best understood as a dispute over the question of what punishment *is*. On one side are "subjectivists" for whom punishment is a means for the production of subjective disutility and for whom the level of suffering inflicted upon individual punishees is crucial (Bronsteen et al. 2009, 2010, 2014; Kolber 2009a,b, 2011, 2012, 2013, 2014). This view is contested by the "objectivists" for whom a punishment for two identical crimes committed by offenders with equal culpability should be nominally equal (Gray 2010; Markel and Flanders 2010).[2,3] The precise level of suffering is also

[2]This debate is a reflection of current developments in neuroscience together with the advances in technologies, such as the functional magnetic resonance imaging (fMRI) or electroencephalogram (EEG), that are starting to allow us to measure individual's perceptions and feelings of pleasure and pain. See, e.g., Brodersen et al. (2012); Civai et al. (2016); Fliessbach et al. (2007); McClure et al. (2004); Schulz et al. (2012); Singer et al. (2004); Tomlin et al. (2006); Weiss et al. (2012). For more general discussion of repercussions of these advances in the realm of the law see Greene and Cohen (2004); Morse (2006, 2011); Miller (2009).

[3]It is also related to the moral philosophy literature which asks whether resources or utility should be equalised to achieve an egalitarian society (Cohen 1989; Dworkin 1981; Sen 1980), for overview see Carter (2004).

incidental for utilitarian theories that emphasise incapacitation or rehabilitation as the primary considerations for sentencing.

These questions are also of interest for deterrence theorists. One may view criminal justice as a monopoly on deterrence using punishment. The standard result in microeconomics that price discrimination increases profits has its analogy in criminal justice. Intuitively, uniform nominal punishment may be inefficient, as some individuals are overdeterred and others underdeterred. The ability to tailor punishment individually should increase efficiency of the system of criminal justice, allowing it to produce more deterrence at lower costs.

This intuition seems to be corroborated by variety models analyzing the relationship between offenders' wealth and optimal sanctions (Arlen 1992; Friedman 1981; Garoupa 1998; Garoupa and Gravelle 2003; Levitt 1997; Lott 1987; Polinsky and Shavell 1984; Polinsky 2006; Polinsky and Shavell 1991). Wealth has been recognized as an important subjective factor in this literature as it determines offenders' ability to pay a fine, and thus her deterrability by a monetary sanction, and affects the severity of any given prison sentence, through the opportunity cost of serving a time in jail. Two frequent predictions of these models stand out: the optimal fines are positively related to income, whereas the optimal (expected) prison sentences and income are related negatively. This result obtains as the optimal fine may exceed the total wealth of some offenders. Prison then supplements fines to achieve deterrence of low-income offenders. These predictions hold when offenders are risk averse, when wealth is unobservable, or when offenders do not have perfect information about the probability of apprehension (Garoupa 1998; Polinsky 2006). Analogous predictions are also generated by models that explicitly incorporate assumptions about risk preferences and severity of sanctions. Risk aversion results in a positive relationship between monetary sanctions and wealth as low-wealth individuals are more easily deterred (Arlen 1992; Polinsky and Shavell 1984). Another important but mostly overlooked source of variability in the severity punishment—and thus in its deterrence as well as retributive value—are lags between crime and punishment.

Listokin (2007) therefore proposes sentencing adjustments in order to maintain discounted sentences constant.

Economists have traditionally argued, that to the extent monetary and nonmonetary sanctions are substitutes, the latter should be only used in cases when the former are ineffective, typically when the optimal fine exceeds offenders' wealth (Becker 1968; Polinsky and Shavell 1984). However, the extent to which different types of sanctions are substitutes is not clearly established (see Kahan 1996; Rizzolli and Tremewan 2016). An additional reason for considering carefully whether punishments should be monetary or non-monetary would arise if people tailor only one type or the other to a punishee's subjective experience, as was found in Montag and Sobek (2014); this would further suggest limited substitutability of different types of sanctions.

3 Experimental Design

3.1 Overview of Experiment

We begin by giving an overview of the experiment and the timing of its individual components, which is summarized in Figure 1.[4] Subjects made decisions in two roles, Role A and Role B. At the beginning, subjects made their Role A decision as to what proportion of their final earnings in the experiment would be donated to a charity. This was followed by three parts which established the variables on which punishments could be conditioned: (i) subjects earned money doing the "Slider Task" (Gill and Prowse 2012), (ii) "Random Payments" were assigned, and (iii) each subject's willingness to pay to avoid repeating the Slider Task (WTP) was elicited.

Next, subjects made their Role B decisions. Each subject was randomly matched with five other subjects in turn, and given the opportunity to assign a punishment for their partner's donation decision. When making each punishment decision, they were given

[4]Full experimental instructions are provided in the Online Appendix.

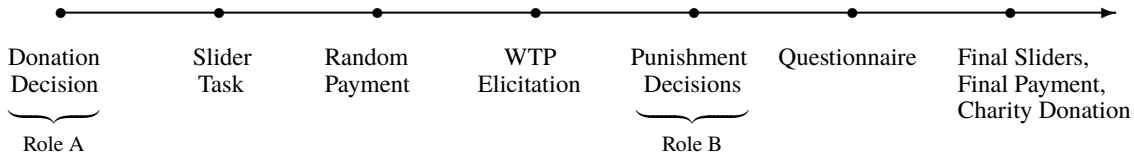

Figure 1: Timeline of the Experiment

information regarding their partner's earnings from the Slider Task, Random Payment, and WTP to avoid the Slider Task.

The type of punishment available depended on the treatment: in the monetary punishment treatment (MP) the participants could impose a fine of up to €5; in the non-monetary punishment treatment (NMP) they could select a number of sliders between 0 and 100 as a punishment. In MP, the fine was subtracted from subjects' earnings in the experiment. In NMP, subjects had to place any punishment sliders they had been assigned before they could collect their payment for the experiment.[5]

After all decisions had been made, subjects were assigned one of the two roles and matched in pairs such that exactly one of each Role B participant's punishment decision was applied to one randomly selected Role A participant. Final payments were then calculated and punishments implemented based on subjects' decisions and role.[6]

3.2 Details of Experimental Tasks

In order that subjects understood the instructions that were to follow, it was necessary for them to be exposed to the Slider Task. On their first screen, subjects were informed that in later parts of the experiment they would be asked to place a number of sliders exactly in

[5]Subjects could, of course, have forgone their payment and left without placing the necessary sliders, however none chose to do so.

[6]We recognize that the subjects' decisions may have been influenced by the fact that they made decisions in both roles. However, this does not affect our estimations of the relationship between punishments and Role A participants' characteristics. This is because Role A participants' characteristics were exogenously assigned to Role B participants through random matching (see Section 4 for details). Similarly, one may be worried that a subject's donation decision may affect their effort in the slider task. However, the determinants of the level of effort Role A subjects exert are not of critical importance to us as the focus of our study is Role B's punishment decisions.

the middle, and shown three sliders that they could move around to get a feel for the task. This screen was shown in both treatments.

The second screen provided subjects with an overview of the experiment. Subjects were informed that there were two roles, Role A and Role B, and that Role A subjects would receive an initial endowment of €10 while Role B subjects would receive €5, that both roles would earn money in three tasks, that their earnings would depend on their effort and luck, and that Role A subjects would make Total Earnings of €10-25 while Role B subjects would make Total Earnings of €5-20.[7] We gave different initial endowments to approximately equalise Final Payments between the roles after donations had been given. Details of the three tasks were not given at this stage.

Subjects were then given a brief description of Role A and Role B decisions, and it was explained that all participants would make decisions in both roles, and that roles (payoff-relevant decisions) would be randomly assigned after all decisions had been made.

Role A Decision: Charity Donation

Subjects were informed that if they were later assigned to Role A, 50% of their Total Earnings would be donated to the Red Cross by default.[8] The donation decision consisted of selecting one of three choices: (i) leave the donation as is, (ii) take back half the donation, or (iii) take back the whole donation. Subjects were asked to make their Role A decisions prior to any earnings to rule out the possibility that those with more wealth could be punished more because they may be seen as having a greater obligation to donate.

We informed the subjects that their actual donation would be transferred online to the Red Cross immediately after payments and the actual donations had been determined at the end of the experiment. In this way, the participants knew in advance that they could see a printed receipt of the payment before they left the laboratory.

[7]For clarity, the initial endowments plus earnings from tasks were always referred to as "Total Earnings" (the amount on which an individual's donation would be based), whereas this amount net of any fine or donation was referred to as the "Final Payment."

[8]We note here that our statistical analysis controls for individual attitudes toward the charity, so the choice of charity is unimportant (see Section 4 for details).

The decision was framed with a 50% donation as the default option in order make smaller donations more likely to appear blameworthy. Clearly, reducing the size of a donation is not a criminal act, but to study our hypotheses we only needed to give subjects a task where some choices would be perceived as being worthy of punishment.[9] We discuss issues of external validity in Section 6.

Slider Task and Bonus

In the Slider Task, subjects had to place 100 sliders precisely in the center of a line using their mouse to drag each slider into the correct position. There were two screens of 50 sliders (see Figure A1 for a screenshot). Each subject could earn a "Bonus", which depended on how quickly they finished relative to other subjects: the first third received €5, the second third €2.50, and the final third received nothing. The size of the bonus received was revealed to each subject as soon as all subjects had finished. Subjects were not informed that this information would be transmitted to the subjects who could punish them, in order to avoid strategic behaviour.

Random Payment

In this part of the experiment, subjects were informed that everyone had the possibility of receiving a "Random Payment": one third would receive €5, another third €2.50, and the remainder nothing. Each subject was then informed of their own random payment.

This random payment was introduced for two main reasons. First, it enables us to see if subjects distinguish between earned and non-earned wealth when deciding punishment levels. Second, we suspected that earnings from the slider task would be strongly correlated with the WTP to avoid the Slider Task, as found in Rizzolli and Tremewan (2016), and

[9] Using a charity as the "victim" rather than implementing a stealing task between subjects, substantially reduced the total number of subjects we had to recruit. The use of a "reverse dictator" game to represent stealing is common in the experimental literature. In Baumann and Friehe (2015); Harbaugh et al. (2013); Khadjavi (2015); Rizzolli and Stanca (2012); and Schildberg-Hörisch and Strassmair (2012), the decision maker may increase their own payoff at the expense of another subject. In Feess et al. (2015), as in our study, it is a charity which suffers the loss.

the Random Payment generates variation in wealth that is independent from subjects' performance in the Slider Task.

Willingness to Pay Elicitation

The elicitation of subjects' willingness to pay to avoid the slider task was performed using a simplified Becker-DeGroot-Marschak method. Subjects were given an extra €5 and told that they could use some or all of this money to avoid repeating the Slider Task. They made 11 binary choices between repeating the Slider Task and paying a sum of money, from €0 - 5 in 50-cent increments. Subjects were informed that one of these choices will be randomly selected to be implemented at the end of the experiment. Choices were forced to be consistent, i.e. if a subject stated they preferred to pay a given amount, they would also have to state that they preferred to pay all smaller amounts. It was clearly stated that if they did not have to repeat the Slider Task, they would not have to wait for others to do so before collecting their payment.[10]

So that subjects understood the concept of WTP when making punishment decisions, it was explicitly stated that the task involved deciding the maximum they were willing to pay to avoid repeating the Slider Task, and that this would be reflected in the highest amount where they chose to pay. As with information on Slider Task Bonuses, in order to avoid strategic behaviour subjects were not informed that their WTP would be transmitted to the subjects who could punish them.

Role B Decisions: Punishments

At this stage, subjects were informed that they would be matched with five other subjects and could decide to impose a punishment if they felt that their partner's choice was not

[10]There is some controversy over the reliability of the Becker-DeGroot-Marschak method (Cason and Plott 2014; Fehr et al. 2015; Plott and Zeiler 2005). However any bias or noise resulting from the mechanism is of no immediate relevance to our results because we do not require a precise measurement of subjects' WTP. What we need is that Role B subjects perceive a Role A subject's higher WTP as a signal of a higher disutility from performing the slider task when determining punishment levels. We view this a reasonable assumption to make and note that this is supported by our main results. A possible bias in WTP measurement may be an issue when comparing the size of punishments across treatments. However, any such bias is unlikely to be sufficient to explain away a difference of the magnitude we observe (see footnote 16).

appropriate. Punishment in our experiment is costless to the punisher, in contrast to most of the literature (but see Feess et al. 2015). The reason for this is that in the environments we are interested in (judges, juries, disciplinary committees), the public actor who makes punishment decisions typically does not bear any costs associated with the punishment.

If at the end of the experiment the punisher was assigned to Role B, one of her five punishment decisions was implemented. The maximum value and type of punishment was explained according to the treatment. In NMP, it was made clear that subjects who did not have to place sliders at the end of the experiment could collect their final payment and leave without having to wait for the others.

Before deciding on each punishment, subjects were informed about their partner's Slider Task Bonus, her Random Payment, and her WTP to avoid the Slider Task. The order in which this information was presented was randomized across subjects. We chose to give all three pieces of information in both treatments in order to minimize experimenter demand effects. Specifically, we were concerned that if we informed subjects only about wealth in MP, and WTP in NMP, and this was all that varied between decisions, they would feel that they were expected to take this information into account and adjust punishments accordingly. While we cannot completely rule out demand effects, the intention of the experimenter should be substantially disguised by having three pieces of information changing between decisions. Thus, our subjects were presented with all available information about the Role A participants and it was up to them to determine, whether, which, and how the individual pieces of information might be relevant for their decisions.

Role B decisions were made using the strategy method: subjects were asked what level of punishment they would assign for each of the three possible donation decisions of their partners; which of their punishment decisions was actually implemented depended on the actual donation decision their partner had made. In MP, they were told that they could choose a fine of up to €5, which would be subtracted from the Role A participant's final

payment, and in NMP they were told they could choose a number of sliders between 0 and 100 which would have to be correctly placed before payment would be received.

Questionnaire

After all payoff-relevant decisions had been made, subjects completed a questionnaire which asked for the subjects gender, age and field of study, as well as responses to the Cognitive Reflection Test (Frederick 2005), and two versions of the trolley dilemma (Edmonds 2014). Finally, subjects were asked questions about whether punishments should vary with wealth, causes of differences in wealth, and their political position on a left-right scale.

Final Sliders and Payment

After completing the questionnaire, subjects were shown the results of the WTP elicitation, which role they had been assigned, and the outcome of the punishment decisions. No subject refused to complete the final sliders, either those resulting from the WTP elicitation or those assigned as a punishment.

3.3 Procedural Details

Three sessions of each treatment were run at the Vienna Center for Experimental Economics. Subjects were recruited via ORSEE (Greiner 2015) and participated in only one session, either MP (74 subjects) or NMP (68 subjects). The experiment was programmed in z-Tree (Fischbacher 2007). Final payments (after donations and fines) ranged from €0.25 to €25.00, with an average of €14.73. A total of €243.21 was donated to the Red Cross.

4 Hypotheses

We are interested in identifying the determinants of the level of punishment meted out by subjects and whether and how these determinants might differ between MP and NMP. We first describe the regression model we use to estimate the importance of different types of information about Role A participants' subjective experience in determining the punishment assigned to them. Then we specify our hypotheses in terms of the parameters of this model.

To identify the determinants of levels of punishment within each treatment, we estimate the following random effects Tobit equation

$$Punishment_{ij} = \beta_S^T STBonus_i + \beta_R^T RandomPayment_i + \beta_W^T WTP_i + \alpha + u_j + v_{ij}, \quad (1)$$

where i identifies Role A subjects (the punishee), and j is the identity of the Role B subject (the punisher), α is the constant, u_j is Role B participant's individual error term (constant across the five punishment decisions), and v_{ij} is the residual. The βs are the coefficients of interest capturing the effect of a Role A participant's subjective characteristics on the punishment assigned to her by the Role B participant, and $T \in MP, NMP$ specifies the treatment. We note that the identifying assumption behind regression (1) is that the Role B subjects' individual error terms, u_js, are not correlated with the right-hand side variables. This requirement is satisfied by construction since Role A and Role B participants are matched randomly in our experiment. That is, Role A participants' characteristics are exogenously assigned to Role B participants.

We emphasise here the fact that our variables of interest are uncorrelated with anything related to the Role B player deciding on the punishment. This means that our main results are unaffected by the aspects of our design that may influence Role B behaviour. Specifically, the fact that subjects have already made a donation decision themselves, how they performed in the Slider Task, and the Random Payment they received will not bias

our estimates of β_S^T, β_R^T, and β_W^T. Nor will any other individual characteristic that may vary across our subject pool.

Equation (1) is estimated separately for each treatment and each donation level. This implicitly controls for heterogeneous attitudes towards the charity. We therefore need not assume that a particular donation decision is blameworthy, it suffices that some subjects see some donation decision as such. However, as the decision was framed with a 50% donation to the Red Cross as the default option, we expect that the subjects will primarily punish the decisions to take all or take half of the donation and the effects of Role A participant characteristics, if any, will be identified in these specifications.

Finally, recall that each participant (while making her Role B decisions) assigned punishment for five other participants' in Role A matched with her. Standard errors are therefore estimated by bootstrapping, accounting for possible correlation of residuals within each subject's Role B decisions.

Our main hypothesis is that subjects take into account the subjective experience of the type of punishment they are inflicting:

Hypothesis 1 *The level of the punishment assigned by a Role B player is decreasing in the disutility of punishment to the Role A player. More specifically:*

A) In MP, fines are increasing in the Role A player's Slider Task Bonus and Random Payment $\left(\beta_S^{MP} > 0 \text{ and } \beta_R^{MP} > 0\right)$.

B) In NMP, the number of sliders assigned as punishment are decreasing in the Role A player's WTP $\left(\beta_W^{NMP} < 0\right)$.

Technically, our WTP elicitation is informative about the subjective trade-off (the marginal rate of substitution) between the disutility from performing the slider task and the utility of the money that must be forgone, rather than being an absolute measure of disutility from performing the slider task. Rather than signaling a high disutility from the slider task, a high WTP might be interpreted as a signal of low utility of money. However, we believed that subjects participating in our experiment would not follow the

latter interpretation for two reasons: (i) It is rather counter-intuitive and requires a degree of economics savvy that is unlikely to be frequent in our subject pool. (ii) Secondly, as subjects were all students who had chosen to earn money in an experiment, there is little reason to suspect a great deal of heterogeneity in their *ex ante* utility of money. In fact, as will become evident, our expectations were born out by the results.

It is possible, however, that subjects resent others' ability or luck, and inflict punishment accordingly, which would lead to a positive correlation between punishment and earnings and between punishment random payment, even in NMP, giving us:

Hypothesis 2 *The level of the punishment assigned by a Role B player is increasing in the Role A player's:*

A) Slider Task Bonus $\left(\beta_S^T > 0\right)$ *and*

B) Random Payment $\left(\beta_R^T > 0\right)$*,*

for $T \in MP, NMP$*.*

Notice that in MP, this hypothesis has an identical prediction to Hypothesis 1A. We will try to disentangle these two hypotheses by comparing the effects of earned and random wealth across treatments: assuming that the effect of resentment is comparable in both treatments, a larger effect in MP would be evidence for an additional effect due to consideration of the Role A player's subjective experience. While we don't have a particular hypothesis in mind, to account for the possibility that there is also some similar channel relating subjective experience of the slider task to punishment in both treatments, we perform a symmetric comparison with WTP.

Hypothesis 3 *The relationship between punishment and the Role A player's disutility of MP (NMP) is stronger in MP (NMP). More specifically:*

A) the relationship between punishment levels and Slider Task Bonus/Random Payment is stronger in MP than in NMP $\left(\beta_S^{MP} > \beta_S^{NMP} \text{ and } \beta_R^{MP} > \beta_R^{NMP}\right)$.

B) the (negative) relationship between punishment levels and WTP is stronger in NMP than in MP $\left(\beta_W^{NMP} < \beta_W^{MP}\right)$.

There is a substantial literature suggesting that people view the entitlement to earned wealth as stronger than the entitlement to windfall gains. This leads to the following hypothesis:

Hypothesis 4 *The (positive) relationship between Slider Task Bonus and punishment levels is weaker than the relationship between Random Payment and punishment levels* $(\beta_S^T > \beta_R^T \; for \; T \in MP, NMP)$.

Finally, we wish to compare the levels of punishment between treatments. This can be done in two ways. First of all, subjects may have an absolute level of disutility in mind, which means that the number of sliders assigned in NMP should be adjusted according to the Role A players WTP before making comparisons with the fines in MP. Alternatively, subjects may see the maximum punishment as a guide to the appropriate punishment for the worst offence, in which case there should be no such adjustment.

Hypothesis 5 *Subjects assign identical punishments in MP and NMP:*

A) in terms of monetary equivalents.

B) as a proportion of the maximum punishment.

5 Results

5.1 Summary Statistics

Table 1 presents the summary statistics of our data set.[11] Subjects took on average approximately 12 minutes to complete the slider task (min: 7 minutes 43 seconds; max: 21 minutes 52 seconds). All possible responses were observed in the WTP elicitation with the median switching-point at €1.50 in both treatments. Mann-Whitney tests[12] find

[11]The z-Tree program, data, and code replicating the results reported in this paper are available at https://dataverse.harvard.edu/dataverse/montag.

[12]Throughout the paper we use Mann-Whitney tests to test for differences in distributions, and stochastic inequality tests to identify directional effects where differences exist (Schlag 2015).

Table 1: Summary Statistics

	Monetary treatment				Non-monetary treatment				Tests (p-values)	
	Mean	25th	Median	75th	Mean	25th	Median	75th	Chi-sq.	Mann-Wh.
Donation decision (%)	19.26	0	25	25	16.91	0	25	25	0.49	0.45
Slider Task Bonus	2.47	0	2.5	5	2.43	0	2.5	5	0.98	0.91
Random Payment	2.47	0	2.5	5	2.43	0	2.5	5	0.98	0.91
Willingness to Pay	1.96	0.5	1.5	3	1.65	0.75	1.5	2.5	0.24	0.63
Total Earnings	16.76	14	15.25	20	16.75	15	17.25	20	0.21	0.82
Final Payment	14.21	12	14	17.5	15.3	12.5	15	17.5	0.12	0.12
Avg. punishment: Take All	2.29	0.4	2	4	42.3	7.5	30	82	–	–
Avg. punishment: Take Half	1.1	0	1	1.8	22.68	0	10.5	41	–	–
Avg. punishment: Do Nothing	0.44	0	0	0.5	12.62	0	0	10	–	–
Time to finish the Slider Task	12:13	10:33	11:56	13:27	12:11	10:31	11:35	13:26	0.6	0.76
Age	25.92	23	25	28	26.4	24	26	28	0.31	0.34
Female (=1)	0.5	0	0.5	1	0.5	0	0.5	1	1	1
Cognitive Reflection Test (0–3)	1.18	0	1	2	1.53	0	2	3	0.2	0.08
Political views (1–5)	2.7	2	3	3	2.68	2	3	3	0.45	0.84
Number of subjects	74				68					

Note: Donation decision is the percentage of Total Earnings donated. Cognitive Reflection Test refers to the number of correct answers. Political views refers to the responses to question "What are your political views in general?": 1 Left; 2 Center Left; 3 Center; 4 Center Right; 5 Right.

no evidence of treatment differences in distributions of donations ($p = 0.45$), completion times ($p = 0.76$), or WTP responses ($p = 0.63$).

For all three possible donation decisions and in both treatments, punishment levels varied from nothing to the maximum possible punishment. The average fine was €0.44 for subjects who left the whole donation to charity, €1.10 for those who took back half, and €2.29 for those who took back everything. For NMP treatment the punishment levels were 13, 23, and 42 sliders respectively. The full distributions of individual subjects' average punishment levels are shown in Figure 2 on page 23. The most notable feature of the distributions are the spikes at zero, which is the modal punishment for each donation decision. Punishment also appears to be constrained by the maximum allowable level, but only for taking the whole donation where maximum punishment is the second most frequent decision.

5.2 Main Results

The results from estimating regression (1) for each donation level and treatment are reported in Table 2. In order to make the estimated coefficients comparable across treatments, we normalize the respective punishment as a percentage of the maximum punishment available.

Table 2: Determinants of Punishment

	Monetary Punishment			Non-Monetary Punishment		
	Take All	Take Half	Take Nothing	Take All	Take Half	Take Nothing
Role A's (punishee) characteristics						
Slider Task Bonus	4.54***	5.46***	9.04*	1.60	1.13*	3.11**
	(1.75)	(1.56)	(4.82)	(1.21)	(0.59)	(1.53)
Random Payment	5.85**	1.78	0.77	1.02	0.89*	1.76
	(2.58)	(1.69)	(5.26)	(0.79)	(0.53)	(1.15)
Willingness to Pay	1.47	3.10	2.25	−3.48**	−1.97**	1.11
	(3.10)	(2.11)	(4.82)	(1.51)	(0.98)	(2.45)
Constant	4.89	−30.55**	−171.26***	38.91***	6.97	−63.52***
	(17.89)	(12.74)	(62.03)	(13.00)	(7.53)	(21.20)
Number of observations	370	370	370	340	340	340
Number of Role B subjects	74	74	74	68	68	68
Random Pay > ST Bonus (z-scores)	0.42	−1.60	−1.16	−0.40	−0.31	−0.71

Note: Double-censored random effects Tobit regressions. The punishments are normalized as the percentage of the maximum available punishment in the respective treatment. Bootstrapped standard errors clustered by Role B participants (punishers) are in parentheses: $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$. At the bottom are reported one-tailed z-tests of the relative importance of the effects of Slider Task Bonus and Random Payment on punishment.

With regard to Hypothesis 1, we find that fines are positively related to both Slider Task Bonus and Random Payment for all three Role A decisions. The coefficients are statistically significant in all cases for Slider Task Bonus, but for Random Payment only when the Role A player takes back the whole donation. In NMP treatment, the number of punishment sliders is decreasing in the Role A subjects WTP, and this relationship is statistically significant when either some or all of the donation is taken.[13]

Result 1 *The level of the punishment assigned by a Role B player is decreasing in the disutility of punishment to the Role A player. More specifically:*

A) In MP, fines are increasing in the Role A player's Slider Task Bonus and Random Payment, and the relationship is statistically significant in four out of six cases.

B) In NMP, the number of sliders assigned as punishment are decreasing in the Role A player's WTP in two out of three cases, and only in these cases is the relationship statistically significant.

Subjects also tend to give larger punishments to those with greater wealth in NMP, where the coefficients on both measures of wealth are also always positive. The coefficient

[13] We have also estimated more flexible specifications with dummies for each level of Slider Task Bonus and Random Payment (€2.5 and €5). The results, reported in Table A1 in the Appendix, are qualitatively similar to our main results from Table 2.

Table 3: Testing Effects Across Treatments

		z-scores		
Coefficients	Test Direction	Take All	Take Half	Take Nothing
Slider Task Bonus	$MP > NMP$	1.38^{*}	2.60^{***}	1.17
Random Payment	$MP > NMP$	1.79^{**}	0.50	−0.18
Willingness to Pay	$NMP < MP$	-1.44^{*}	-2.18^{**}	−0.21

Note: One-tailed z-test computed using the coefficients and standard errors reported in Table 2: $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

on Slider Task Bonus is significant when the Role A subject takes back nothing, and both coefficients are weakly significant when they take back half. This is consistent with our subjects assigning punishment out of spite (Hypothesis 2), rather than accounting for the effect of punishment on the punishee's subjective experience (Hypothesis 1A).

Result 2 *In both MP and NMP, the level of the punishment assigned by a Role B player is increasing in the Role A player's:*

A) Slider Task Bonus, statistically significant in five of six cases, and

B) Random Payment, statistically significant in two of six cases.

In order to disentangle these two hypotheses, we compare the magnitude of coefficients across our two treatments (Hypothesis 3). The results are reported in Table 3. The effect of earned and random wealth is greater in MP for every comparison except for the effect of random payment when Role A takes nothing, where the difference is small in magnitude and not statistically significant. The difference is statistically significant in three of the five remaining cases, where the direction is in line with our hypotheses.

These results suggest that the positive relationship between wealth and monetary punishment is not entirely driven by resentment of those who do well. Rather, there is a stronger tendency to increase monetary punishment with Role A's wealth when the punishment is monetary in nature, giving support to Hypothesis 1.

Finally, we also note that WTP was never significant in MP and that the effect on punishment is significantly larger (in the absolute terms) in NMP when the Role A participant took back all or half of the donation. In sum, when the information is relevant for

subjective experience of the respective punishment, it has a stronger impact on punishment decisions than when it is actually not relevant. We interpret this as evidence of our main hypotheses even if other motives for conditioning on these variables exist.

Result 3 *The relationship between punishment and the Role A player's disutility of MP (NMP) is typically stronger in MP (NMP). More specifically:*

A) in five out of six cases the relationship between punishment levels and Slider Task Bonus/Random Payment is stronger in MP than in NMP, and in three of these the difference is statistically significant.

B) the relationship between punishment levels and WTP is stronger in NMP than in MP, and in two of three cases the difference is statistically significant.

With regard to Hypothesis 4, the test of the relative effect of random and earned wealth on punishment are reported at the bottom of Table 2. We do not find evidence that random wealth is a stronger determinant of punishment than earned wealth. If anything, the signs of the z-scores are mostly negative, suggesting a larger role of earned wealth.

We summarize the specific results pertaining to our Hypotheses 1 to 4 set up in Section 4 as follows:

Result 4 *We find no evidence that punishments are more sensitive to random wealth than earned wealth.*

We now turn to testing for treatment differences in the levels of punishments (Hypothesis 5), first comparing the monetary equivalents, then proportions of the maximum possible punishment. Because our non-parametric statistical tests will require independent observations, we use the average punishment chosen by each subject.[14] To make our first comparison across treatments, we wish to replace the number of sliders given as punishment with the equivalent monetary value as experienced by the Role A subject. In

[14]Note that because the distributions of Slider Task bonus and Random Payment are identical across treatments by construction and neither the distributions of WTP, nor subjects' demographic characteristics were found to differ across the treatments (see Table 1), there is no need to control for these factors in this analysis.

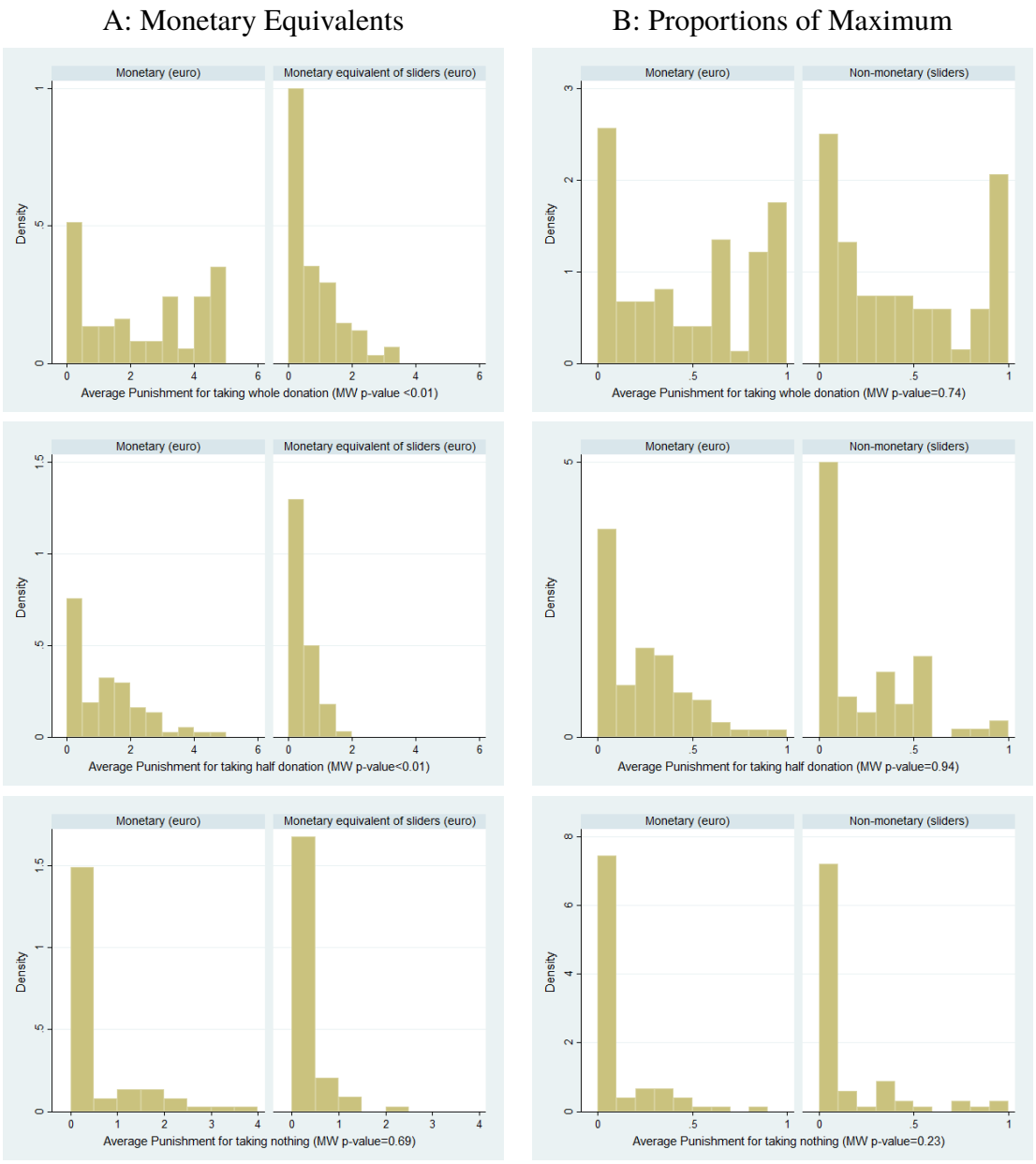

Figure 2: Distributions of average punishments. Panel A: slider punishment is represented in monetary terms using the elicited WTP to compute the monetary equivalent. Panel B: punishments are expressed as proportions of the maximum available punishment.

order to approximate this we assume that utility is linear in money, and that the true WTP of the subject is the midpoint of the interval we have elicited.[15]

The resulting histograms are shown in Panel A of Figure 2. For taking back all or half of the donation, the mass of the distribution of monetary equivalents of the slider punishments appears to be shifted substantially towards lower punishment. Indeed, the average punishments are roughly one third the size in NMP, and stochastic inequality tests show that in both cases non-monetary punishments tend to be smaller than fines ($p < 0.01$).[16] There is no evidence that the distributions are different when the Role A subject took back nothing (Mann-Whitney, $p = 0.69$).

Panel B of Figure 2 show punishment levels as a proportion of the maximum objective payoff (€5 or 100 sliders). Here the distributions look very similar across treatments, and Mann-Whitney tests find no statistically significant difference (Take all: $p = 0.74$; Take half: $p = 0.94$; Take nothing: $p = 0.23$).[17]

Result 5 *Comparison of punishment levels in MP and NMP:*

A) Subjects assign higher punishments in MP than NMP in terms of monetary equivalents.

B) We find no evidence that subjects assign different punishments in MP and NMP as a proportion of the maximum punishment.

[15] If subjects stated that they were willing to pay every amount, then all we know is that there WTP is greater than €5. We used the same formula for these subjects as the others and assumed their WTP was €5.25. The reader may question our transformation of non-monetary punishments into monetary equivalents by noting that utility need not be linear in money. At the same time, we note that the disutility from the Slider Task is also likely to be non-linear. In order to be able to do the transformation exactly, we would have to do multiple WTP elicitations for different numbers of sliders, which was impractical within our experimental design.

[16] These results are also consistent with a downward bias in elicited WTP caused by the Becker-DeGroot-Marschak mechanism, however the size of the difference is much greater that any WTP-WTA gap observed in the literature discussed in footnote 10.

[17] Regressions controlling for Role B characteristics also find significantly lower punishment in NMP when measured in monetary equivalents, and no difference in punishment levels when measured as a proportion of maximum punishment (see Table A2 in the Appendix).

5.3 Additional Results

Having dealt with the core hypotheses of our paper, we now further exploit the information collected during the experiment in order to study the possible roles of other factors that may have a bearing on punishment levels.

As in many of the studies cited in Section 2, we find that greater deviations from a social norm are met with harsher punishment. In our experiment the most appropriate action is not necessarily clear: some may consider keeping money they earned as entirely moral. Nevertheless, we think it reasonable to assume that taking more from the charity is seen as at least as bad as taking less. Under this assumption we can look at the size of deviations from a social norm in two ways: firstly we consider increases in the amount Role A player takes; secondly we consider changes in the amount Role B participants would have taken, viewing this as an indication of their perception of an appropriate action.

Visually comparing the distributions of punishment across donation decisions within each treatment in Figure 2, one can see that punishments do indeed tend to be higher the more that Role A participant takes from the charity. Sign tests confirm in both treatments that those who take everything are punished more harshly than those who take half ($p < 0.01$), who are in turn punished more harshly than those who take nothing ($p < 0.01$). This result can also be seen from the regressions in Table 2, where for both treatments the constants are monotonically increasing in the amount taken. The differences in constants reported in Table 2 are statistically significant at the 5% level for all comparisons apart from the difference between taking everything and taking half in the monetary treatment. In summary, these results suggest that our subjects tended to consider taking back the charity donation as blameworthy.

Table 4 reports results of estimating regressions (1) together with Role B (punisher) participants' own decisions and characteristics. First, we note that all the coefficients of interest, that is Role A participant's characteristics, are numerically as well as statistically almost identical to our main estimates in Table 2. This is, is not surprising since participants are matched randomly, so that Role A participants' characteristics are exogenous for Role

Table 4: Determinants of Punishment: Controlling for Role B Decisions and Characteristics

	Monetary Punishment			Non-Monetary Punishment		
	Take All	Take Half	Take Nothing	Take All	Take Half	Take Nothing
Role A's (punishee) characteristics						
Slider Task Bonus	4.43**	5.33***	8.83*	1.67	1.17**	3.14**
	(1.80)	(1.52)	(4.87)	(1.21)	(0.59)	(1.56)
Random Payment	5.75**	1.73	0.57	0.98	0.86	1.73
	(2.57)	(1.67)	(5.30)	(0.79)	(0.53)	(1.14)
Willingness to Pay	1.40	3.10	2.47	−3.61**	−2.02**	1.15
	(3.13)	(2.10)	(5.05)	(1.54)	(0.99)	(2.51)
Role B's (punisher) characteristics						
Slider Task Bonus	−1.99	−7.72*	−14.95	0.95	1.22	−0.29
	(7.52)	(4.28)	(27.09)	(4.90)	(3.10)	(6.05)
Random Payment	−2.79	1.39	0.83	−3.39	−1.41	2.57
	(6.22)	(3.50)	(15.28)	(5.00)	(2.86)	(5.55)
Willingness to Pay	8.01	2.85	−8.56	3.90	4.62	17.62
	(7.95)	(4.34)	(13.89)	(8.25)	(5.18)	(11.06)
Donation decision (%)	3.07***	0.31	−0.95	2.35***	0.87**	0.91
	(1.01)	(0.52)	(2.77)	(0.78)	(0.41)	(0.95)
Age	−2.52	−0.91	3.75	2.38	0.46	−3.95
	(3.06)	(1.70)	(11.42)	(2.16)	(1.22)	(3.23)
Female (=1)	−16.24	−17.98	−47.17	22.87	11.64	35.42
	(30.30)	(17.07)	(61.87)	(25.59)	(14.93)	(26.71)
Cognitive Reflection Test (0–3)	−3.88	−11.19	−28.37	0.09	−2.33	−3.09
	(14.15)	(8.41)	(43.66)	(9.23)	(5.23)	(9.32)
Political views (0–5)	−9.12	−3.06	1.21	21.76**	12.58**	18.98*
	(14.56)	(8.85)	(41.83)	(9.65)	(5.16)	(10.94)
Constant	44.71	28.30	−141.91	−134.38*	−63.49*	−71.53
	(126.55)	(65.35)	(384.83)	(68.89)	(38.12)	(84.78)
Number of observations	370	370	370	340	340	340
Number of Role B subjects	74	74	74	68	68	68

Note: Double-censored random effects Tobit regressions. The punishments are normalized as the percentage of the maximum available punishment in the respective treatment. Donation decision is the percentage of Total Earnings donated. Cognitive Reflection Test refers to the number of correct answers. Political views refers to the responses to question "What are your political views in general?": 1 Left; 2 Center Left; 3 Center; 4 Center Right; 5 Right. Bootstrapped standard errors clustered by Role B participants (punishers) are in parentheses: $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

B participants, but it is reassuring if one would be worried about a possible random correlation between Role A participants and Role B participants' characteristics.

The results suggest that Role B participants own donation decision predicts the size of punishment for taking back the whole donation and taking half of the donation in the non-monetary treatment. This is consistent the expectation that subjects who take less perceiving taking from the charity as worse than those who take more. Somewhat surprisingly, own donation decision does not seem to predict punishment for taking half in the monetary treatment. We also find that more right wing political views are related

to harsher punishment in the non-monetary treatment, but not in the monetary treatment. None of the other Role B players' characteristics, including their own Slider Task Bonus, Random Payment, and WTP, are systematically related to punishment in our data.

To further investigate how differences in individual norms may play a role in punishment behaviour, we divided subjects into three categories: (i) those who within all three donation decisions assigned the same (positive) punishment to all five Role A participants, (ii) those who gave different punishments to different Role A players, and (iii) those who never punished. We categorized those who never punished separately from uniform punishment because we regard these as people for whom the threshold for punishment had not been met, and so we do not know whether they would actually have varied punishment levels if the threshold had been met.

Using the unconditional exact z-tests (Suissa and Shuster 1985) we find that: Those who take everything choose disproportionately to never punish, 41% compared with 5% of those who took back half ($p < 0.01$) and 7% of those who took back nothing ($p = 0.019$). This is consistent with those who took everything being less likely to perceive doing so as a deviation from the norm, and so not worthy of punishment. Those who took back half choose disproportionately to vary their punishments: 85% compared with 48% of those who take everything ($p < 0.01$) and 53% of those who took nothing ($p = 0.011$). Those who took back nothing choose disproportionately punish uniformly: 40% compared with 11% of those who take everything ($p = 0.019$) and 10% of those who took back half ($p = 0.010$).

We can also look at individual behaviour to identify spiteful punishment. We take a conservative approach and categorize an individual as spiteful if they assign positive punishment in all decisions. Assigning positive punishment for all possible actions of Role A implies that the punishment cannot be purely for deviation from a perceived norm, regardless of the norm held by the punisher. We find significantly more spiteful individuals

in NMP (22%) than MP (7%).[18] Interestingly, the proportion of such individuals is almost identical across all three donation levels (between 13% and 15%).

6 Discussion and Conclusions

We ran a laboratory experiment to study whether or not people find it acceptable to give different nominal levels of punishment to different individuals for the same offence. We find that subjects do condition levels of punishment on information about the punishee's likely subjective experience of the punishment, and that this is true for both monetary and non-monetary punishments. In addition, we find that subjects are significantly influenced by the maximum nominal punishment available when determining the precise level of punishment, rather than having in mind a particular quantity of disutility they would like to inflict. In line with earlier studies, we find that greater deviation from the social norm is met with harsher punishment, and that some subjects appear to punish for spiteful reasons.

In our experiment, the subjects are students, the act which is punished is not a real crime, and the non-monetary punishment which can be imposed does not share many of the features of the kinds of punishments that are used in practice. What, if anything, can this laboratory experiment tell us about our domains of interest: punishment decisions made by judges, juries, and other public actors?

It is helpful here to make the distinction between "ecological" and "external validity" (Frechette 2015). Ecological validity relates to the degree to which tasks in the experiment resemble tasks in the field; external validity relates to the degree to which conclusions from an experiment are likely to hold in the field. These concepts are quite different, and neither one implies the other. For example, the results of an experiment which precisely replicates a real world environment, and thus has 100% ecological validity, may be driven by the fact that subjects feel they are observed by the experimenters, and fail to transfer

[18] This is not simply a reflection of the minimum positive punisment being smaller in NMP - relaxing our definition such that subjects in NMP are categorized as spiteful only if their minimum punishment is at least 10% of the maximum (the minimum proportion possible in MP) yields similar results.

outside the laboratory. On the other hand, the abstract games played in typical economic experiments have no counterpart in everyday life, but are widely held to reveal preferences towards such things as fairness and reciprocity, and yield insights into a variety of real life decision making scenarios, such as tax compliance and how to best incentivize workers.

Clearly the present experiment has limited ecological validity, however we believe that our results have a significant degree of external validity. In our experiment, we find that subjects who consider a particular act as worthy of punishment, adjust the level of punishment they impose according to information about the punishee's subjective experience of that punishment. We see this pattern as resulting from underlying views about fairness and justice, and see no *a priori* reason why our results should not extend to other acts deemed worthy of punishment (e.g. crimes) and other forms of punishment (e.g. imprisonment). Of course, institutional factors and salient features of particular forms of punishment not present in our experimental environment may limit the importance of our findings to the actual decisions of judges and juries. However, the basic insights our results provide into people's attitudes towards subjective determinants of punishment are still relevant.[19]

The implications of our study depend on one's view on the nature and purpose of punishment. Our results are good news for those who believe that "the punishment should fit the crime" and equate punishment with subjective suffering, and for those who argue that punishment needs to be tailored individually to improve deterrence and increase efficiency of law enforcement expenditures: such policies may be not be excessively controversial. For those who hold that justice requires equal nominal punishment for equivalent crimes, our results may be disturbing. We find a clear tendency to adjust punishments according to signals regarding a punishee's subjective experience, which is likely to extend to jury, and possibly judge, decision-making. To counteract this tendency, those with an objectivist view of justice should push for sentencing guidelines and restrictions on judicial discretion.

[19]We also note that there is evidence that student subjects behave much like the rest of the population in economic experiments (Frechette 2015).

In our experiment, punishments in MP were harsher than the monetary equivalents of punishments in NMP. One might be inclined to interpret this as a socially desirable outcome, since non-monetary punishments are socially costly (in our case a pure waste of time and effort), whereas monetary punishments are just transfers with minimal social cost. This might be an alternative explanation of the differences between our two treatments seen in Figure 2. However, the sameness of distributions of both types of punishment when computed as a proportion of the maximum punishment, rather suggests that it was the range of available sentences that guided our subjects. Disentangling these two hypotheses rigorously is an important question for future research, as the exact role of sentencing limits has significant policy implications. Namely, statutory limitations may shift the entire distribution of punishment decisions, rather than achieve the desired goal of simply ensuring a minimum or maximum sentence for a particular crime.

Finally, when comparing our incentivised laboratory data with the earlier vignette study raises a number of other interesting questions. For monetary punishment, both our and Montag and Sobek's (2014) earlier findings consistently show that individuals do vary the level of punishment with the subjective effect of that punishment on the person to be punished. Moreover, in our post-experimental questionnaire, 50% of our subjects responded that fines for a given offence should be increasing in wealth.

However, the experimental results for non-monetary punishment are mixed. Subjects in Montag and Sobek's (2014) study, who were explicitly asked to decide on hypothetical prison sentences for convicted defendants, exhibited a strong tendency to impose nominally uniform punishment. There is evidence of the same attitude regarding incarceration in our subject pool where 75% responded in the post-experimental questionnaire that prison sentences should not be conditioned on wealth. There is, in fact, no discernible relationship between our subjects' questionnaire responses and their punishment decisions: subjects who responded that sentences should not depend on wealth were not more likely to punish uniformly.

Our study shows that peoples' unwillingness to vary prison sentences is unlikely to be due purely to their non-monetary nature. Further research is therefore necessary to ascertain the cause of this inconsistency. Unincentivised tasks and survey questions regarding morally charged issues may fail to elicit true preferences. Alternatively, imprisonment may differ in some fundamental way from the non-monetary punishment we use in our experiment, such as through the role of social stigma, which was absent in our anonymous environment. The answer to this puzzle may have important repercussions for economists', policy makers', as well as lawyers' understanding of criminal punishment and what policies they should recommend.

References

Abbink, Klaus and Benedikt Herrmann. 2011. "The moral costs of nastiness." *Economic Inquiry* 49:631–633.

Alatas, Vivi, Lisa Cameron, Ananish Chaudhuri, Nisvan Erkal, and Lata Gangadharan. 2009. "Gender, culture, and corruption: Insights from an experimental analysis." *Southern Economic Journal* 75:663–680.

Alm, James. 2012. "Measuring, explaining, and controlling tax evasion: lessons from theory, experiments, and field studies." *International Tax and Public Finance* 19:54–77.

Arlen, Jennifer H. 1992. "Should Defendants' Wealth Matter?" *Journal of Legal Studies* 21:413–429.

Baumann, Florian and Tim Friehe. 2015. "Proof beyond a reasonable doubt: Laboratory evidence." DICE Discussion Paper No. 181, Düsseldorf Institute for Competition Economics, Heinrich-Heine University Düsseldorf.

Becker, Gary S. 1968. "Crime and Punishment: An Economic Approach." *Journal of Political Economy* 76:169–217.

Bernhard, Helen, Ernst Fehr, and Urs Fischbacher. 2006. "Group affiliation and altruistic norm enforcement." *American Economic Review* 96:217–221.

Brodersen, Kay H., Katja Wiech, Ekaterina I. Lomakina, Chia-shu Lin, Joachim M. Buhmann, Ulrike Bingel, Markus Ploner, Klaas Enno Stephan, and Irene Tracey. 2012. "Decoding the Perception of Pain from fMRI Using Multivariate Pattern Analysis." *NeuroImage* 63:1162–1170.

Bronsteen, John, Christopher Buccafusco, and Jonathan Masur. 2009. "Happiness and Punishment." *University of Chicago Law Review* 76:1037–1082.

———. 2010. "Retribution and the Experience of Punishment." *California Law Review* 98:1463–1496.

———. 2014. *Happiness and the Law*. Chicago: University of Chicago Press.

Carpenter, Jeffrey and Peter Hans Matthews. 2009. "What norms trigger punishment?" *Experimental Economics* 12:272–288.

———. 2012. "Norm enforcement: anger, indignation, or reciprocity?" *Journal of the European Economic Association* 10:555–572.

Carter, Alan. 2004. "The Quest for an Egalitarian Metric." *Critical Review of International Social and Political Philosophy* 7:94–113.

Cason, Timothy N. and Charles R. Plott. 2014. "Misconceptions and Game form Recognition: Challenges to Theories of Revealed Preference and Framing." *Journal of Political Economy* 122:1235–1270.

Charness, Gary, Ramón Cobo-Reyes, and Natalia Jiménez. 2008. "An investment game with third-party intervention." *Journal of Economic Behavior & Organization* 68:18–28.

Civai, Claudia, Daniel R. Hawes, Colin G. DeYoung, and Aldo Rustichini. 2016. "Intelligence and Extraversion in the Neural Evaluation of Delayed Rewards." *Journal of Research in Personality* 61:99–108.

Cohen, Gerald A. 1989. "On the Currency of Egalitarian Justice." *Ethics* 99:906–944.

DeVeaux, Mika'il. 2013. "The Trauma of the Incarceration Experience." *Harvard Civil Rights-Civil Liberties Law Review* 48:257–277.

Donna, Javier and José-Antonio Espín-Sánchez. 2015. "Let the Punishment Fit the Criminal." Working paper, Department of Economics, Ohio State University, Columbus, OH.

Dworkin, Ronald. 1981. "What is Equality? Part 2: Equality of Resources." *Philosophy & Public Affairs* 10:283–345.

Edmonds, David. 2014. *Would You Kill the Fat Man? The Trolley Problem and What Your Answer Tells Us about Right and Wrong*. Princeton: Princeton University Press.

Edney, Richard. 2004. "To Keep Me Safe from Harm? Transgender Prisoners and the Experience of Imprisonment." *Deakin Law Review* 9:327–338.

Fazel, Seena and Jacques Baillargeon. 2011. "The Health of Prisoners." *The Lancet* 377:956–965.

Feess, Eberhard, Hannah Schildberg-Hörisch, Markus Schramm, and Ansgar Wohlschlegel. 2015. "The impact of fine size and uncertainty on punishment and deterrence: Evidence from the laboratory." IZA Discussion Paper No. 9388, Institute for the Study of Labor, Bonn.

Fehr, Dietmar, Rustamdjan Hakimov, and Dorothea Kübler. 2015. "The Willingness to Pay–Willingness to Accept Gap: A Failed Replication of Plott and Zeiler." *European Economic Review* 78:120–128.

Fehr, Ernst and Urs Fischbacher. 2004. "Third-party punishment and social norms." *Evolution and Human Behavior* 25:63–87.

Fišar, Miloš, Matúš Kubák, Jiři Špalek, and James Tremewan. 2016. "Gender differences in beliefs and actions in a framed corruption experiment." *Journal of Behavioral and Experimental Economics* 63:69–82.

Fischbacher, Urs. 2007. "z-Tree: Zurich Toolbox for Ready-Made Economic Experiments." *Experimental Economics* 10:171–178.

Fliessbach, K., B. Weber, P. Trautner, T. Dohmen, U. Sunde, C. E. Elger, and A. Falk. 2007. "Social Comparison Affects Reward-Related Brain Activity in the Human Ventral Striatum." *Science* 318:1305–1308.

Frechette, Guillaume R. 2015. "Laboratory Experiments: Professionals versus Students." In *Handbook of Experimental Economic Methodology*, edited by Guillaume R. Frechette and Andrew Schotter. Oxford University Press, 360–390.

Frederick, Shane. 2005. "Cognitive Reflection and Decision Making." *Journal of Economic Perspectives* 19:25–42.

Friedman, David D. 1981. "Reflections on optimal punishment, or: Should the rich pay higher fines?" *Research in Law and Economics* 3:185–205.

Garoupa, Nuno. 1998. "Optimal Law Enforcement and Imperfect Information When Wealth Varies among Individuals." *Economica* 65:479–490.

Garoupa, Nuno and Hugh Gravelle. 2003. "Efficient Deterrence does not Require that the Wealthy should be Able to Buy Justice." *Journal of Institutional and Theoretical Economics* 159:545–552.

Gill, David and Victoria Prowse. 2012. "A Structural Analysis of Disappointment Aversion in a Real Effort Competition." *American Economic Review* 102:469–503.

Granic, Dura-Georg and Alexander K. Wagner. 2017. "Tie-Breaking Power in Committees." Working paper, Erasmus School of Economics, Erasmus University Rotterdam.

Gray, David. 2010. "Punishment as Suffering." *Vanderbilt Law Review* 63:1619–1693.

Greene, Joshua and Jonathan Cohen. 2004. "For the Law, Neuroscience Changes Nothing and Everything." *Philosophical Transactions of the Royal Society of London B: Biological Sciences* 359:1775–1785.

Greiner, Ben. 2015. "Subject Pool Recruitment Procedures: Organizing Experiments with ORSEE." *Journal of the Economic Science Association* 1:114–125.

Harbaugh, William T., Naci Mocan, and Michael S. Visser. 2013. "Theft and deterrence." *Journal of Labor Research* 34:389–407.

Henrich, Joseph, Richard McElreath, Abigail Barr, Jean Ensminger, Clark Barrett, Alexander Bolyanatz, Juan Camilo Cardenas, Michael Gurven, Edwins Gwako, Natalie Henrich, Carolyn Lesorogol, Frank Marlowe, David Tracer, and John Ziker. 2006. "Costly Punishment Across Human Societies." *Science* 312:1767–1770.

Herrmann, Benedikt, Christian Thöni, and Simon Gächter. 2008. "Antisocial punishment across societies." *Science* 319:1362–1367.

Hoff, Karla, Mayuresh Kshetramade, and Ernst Fehr. 2011. "Caste and punishment: the legacy of caste culture in norm enforcement." *Economic Journal* 121:F449–F475.

Kahan, Dan M. 1996. "What do Alternative Sanctions Mean?" *University of Chicago Law Review* 63:591–653.

Khadjavi, Menusch. 2015. "On the interaction of deterrence and emotions." *Journal of Law, Economics, and Organization* 31:287–319.

Kolber, Adam J. 2009a. "The Comparative Nature of Punishment." *Boston University Law Review* 89:1565–1608.

———. 2009b. "The Subjective Experience of Punishment." *Columbia Law Review* 109:182–236.

———. 2011. "The Experiential Future of the Law." *Emory Law Journal* 60:585–652.

———. 2012. "Unintentional Punishment." *Legal Theory* 18:1–29.

———. 2013. "Against Proportional Punishment." *Vanderbilt Law Review* 66:1141–1309.

———. 2014. "Will There Be a Neurolaw Revolution." *Indiana Law Journal* 89:807.

Leibbrandt, Andreas and Raúl López-Pérez. 2012. "An exploration of third and second party punishment in ten simple games." *Journal of Economic Behavior & Organization* 84:753–766.

Levitt, Steven D. 1997. "Incentive Compatibility Constraints as an Explanation for the Use of Prison Sentences Instead of Fines." *International Review of Law and Economics* 17:179–192.

Listokin, Yair. 2007. "Crime and (with a lag) Punishment: The Implications of Discounting for Equitable Sentencing." *American Criminal Law Review* 44:115–140.

Lott, John R. 1987. "Should the Wealthy Be Able to "Buy Justice"?" *Journal of Political Economy* 95:1307–1316.

———. 1992. "Do We Punish High Income Criminals Too Heavily?" *Economic Inquiry* 30:583–608.

Markel, Dan and Chad Flanders. 2010. "Bentham on Stilts: The Bare Relevance of Subjectivity to Retributive Justice." *California Law Review* 98:907–988.

Massoglia, Michael. 2008. "Incarceration as Exposure: The Prison, Infectious Disease, and Other Stress-Related Illnesses." *Journal of Health and Social Behavior* 49:56–71.

Massoglia, Michael, Brianna Remster, and Ryan D. King. 2011. "Stigma or Separation? Understanding the Incarceration-Divorce Relationship." *Social Forces* 90:133–155.

McClure, Samuel M., David I. Laibson, George Loewenstein, and Jonathan D. Cohen. 2004. "Separate Neural Systems Value Immediate and Delayed Monetary Rewards." *Science* 306:503–507.

Miller, Greg. 2009. "Brain Scans of Pain Raise Questions for the Law." *Science* 323:195.

Montag, Josef and Tomáš Sobek. 2014. "Should Paris Hilton Receive a Lighter Prison Sentence Because She's Rich: An Experimental Study." *Kentucky Law Journal* 103:95–125.

Morse, Stephen J. 2006. "Brain Overclaim Syndrome and Criminal Responsibility: A Diagnostic Note." *Ohio State Journal of Criminal Law* 3:397–412.

———. 2011. "Neuroscience and the Future of Personhood and Responsibility." In *Constitution 3.0: Freedom and Technological Change*, edited by Jeffrey Rosen and Benjamin Wittes. Washington, D.C: Brookings Institution Press, 113–129.

Nelissen, Rob M. A. and Marcel Zeelenberg. 2009. "Moral emotions as determinants of third-party punishment: Anger, guilt and the functions of altruistic sanctions." *Judgment and Decision making* 4:543.

Ohtsubo, Yohsuke, Fumiko Masuda, Esuka Watanabe, and Ayumi Masuchi. 2010. "Dishonesty invites costly third-party punishment." *Evolution and Human Behavior* 31:259–264.

Pager, Devah. 2003. "The Mark of a Criminal Record." *American Journal of Sociology* 108:937–975.

Plott, Charles R. and Kathryn Zeiler. 2005. "The Willingness to Pay–Willingness to Accept Gap, the." *American Economic Review* 95:530–545.

Polinsky, A. Mitchell. 2006. "The Optimal Use of Fines and Imprisonment when Wealth is Unobservable." *Journal of Public Economics* 90:823–835.

Polinsky, A. Mitchell and Steven Shavell. 1984. "The Optimal Use of Fines and Imprisonment." *Journal of Public Economics* 24:89–99.

———. 1991. "A Note on Optimal Fines When Wealth Varies among Individuals." *American Economic Review* 81:618–621.

Posner, Richard A. 1985. "An Economic Theory of the Criminal Law." *Columbia Law Review* 85:1193–1231.

Raaijmakers, Ellen A. C., Thomas A. Loughran, Jan W. de Keijser, Paul Nieuwbeerta, and Anja J. E. Dirkzwager. 2017. "Exploring the Relationship between Subjectively Experienced Severity of Imprisonment and Recidivism: A Neglected Element in Testing Deterrence Theory." *Journal of Research in Crime and Delinquency* 54:3–28.

Rizzolli, Matteo and Luca Stanca. 2012. "Judicial errors and crime deterrence: theory and experimental evidence." *Journal of Law and Economics* 55:311–338.

Rizzolli, Matteo and James Tremewan. 2016. "Hard Labour in the Lab: Are Monetary and Non-Monetary Sanctions Really Substitutable?" Working Paper No. 1606, Department of Economics, University of Vienna.

Schildberg-Hörisch, Hannah and Christina Strassmair. 2012. "An experimental test of the deterrence hypothesis." *Journal of Law, Economics, and Organization* 28:447–459.

Schlag, Karl H. 2015. "Who gives Direction to Statistical Testing? Best Practice meets Mathematically Correct Tests." Working Paper No. 1512, Department of Economics, University of Vienna.

Schulz, Enrico, Andrew Zherdin, Laura Tiemann, Claudia Plant, and Markus Ploner. 2012. "Decoding an Individual's Sensitivity to Pain from the Multivariate Analysis of EEG Data." *Cerebral Cortex* 22:1118–1123.

Sen, Amartya. 1980. "Equality of What?" In *Tanner Lectures on Human Values*, vol. I, edited by Sterling M. McMurrin. Cambridge, UK: Cambridge University Press, 195–220.

Singer, Tania, Ben Seymour, John O'Doherty, Holger Kaube, Raymond J. Dolan, and Chris D. Frith. 2004. "Empathy for Pain Involves the Affective but not Sensory Components of Pain." *Science* 303:1157–1162.

Suissa, Samy and Jonathan J Shuster. 1985. "Exact unconditional sample sizes for the 2×2 binomial trial." *Journal of the Royal Statistical Society. Series A (General)* 148:317–327.

Tomlin, Damon, M. Amin Kayali, Brooks King-Casas, Cedric Anen, Colin F. Camerer, Steven R. Quartz, and P. Read Montague. 2006. "Agent-Specific Responses in the Cingulate Cortex During Economic Exchanges." *Science* 312:1047–1050.

Weiss, Alexander, James E. King, Miho Inoue-Murayama, Tetsuro Matsuzawa, and Andrew J. Oswald. 2012. "Evidence for a Midlife Crisis in Great Apes Consistent with the U-shape in Human Well-being." *Proceedings of the National Academy of Sciences* 109:19949–19952.

Williams, Hannah Kate, Pamela J. Taylor, Julian Walker, Gemma Plant, Anna Kissell, and Amy Hammond. 2013. "Subjective Experience of Early Imprisonment." *International Journal of Law and Psychiatry* 36:241–249.

Appendix

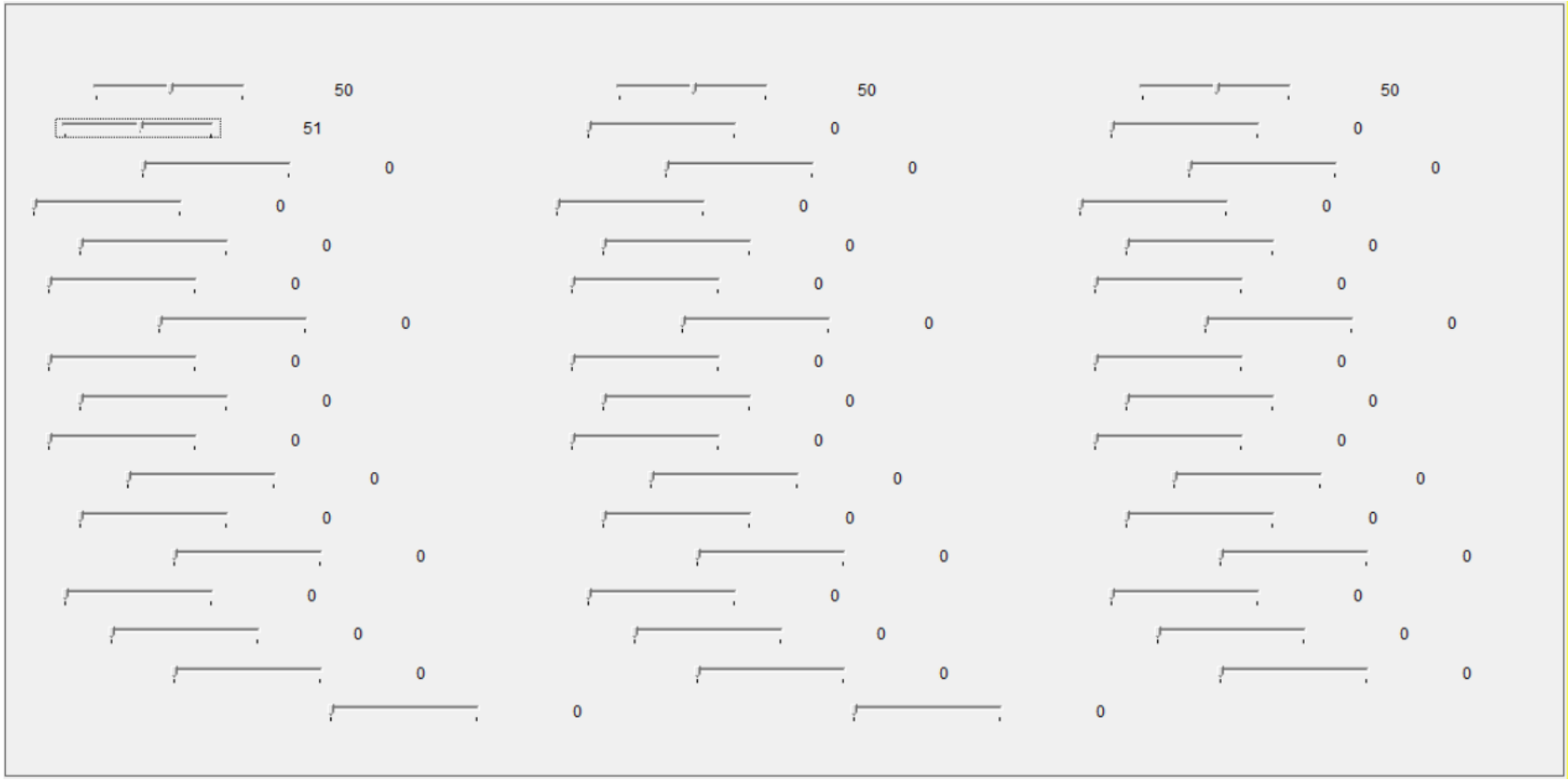

Figure A1: Slider task (one half) screen shot, 50 sliders

Table A1: Determinants of Punishment: Flexible Specification

	Monetary Punishment			Non-Monetary Punishment		
	Take All	Take Half	Take Nothing	Take All	Take Half	Take Nothing
Role A's (punishee) characteristics						
Slider Task Bonus = €2.5	−0.86	14.24**	20.43	9.29*	5.14*	2.42
	(11.28)	(6.27)	(21.88)	(5.48)	(3.03)	(5.54)
Slider Task Bonus = €5	21.59**	30.48***	46.31*	7.94	5.66*	15.02**
	(8.96)	(8.07)	(25.00)	(5.95)	(2.97)	(7.34)
Random Payment = €2.5	23.62**	−7.75	−2.31	0.99	1.46	−0.35
	(10.57)	(7.05)	(17.57)	(4.07)	(2.51)	(5.62)
Random Payment = €5	29.50**	9.40	4.42	5.04	4.47*	7.92
	(13.00)	(8.12)	(26.31)	(3.92)	(2.68)	(5.15)
Willingness to Pay	1.33	3.79*	2.58	−3.26**	−1.83	0.36
	(3.17)	(2.19)	(5.39)	(1.56)	(1.16)	(2.74)
Constant	6.67	−29.35**	−170.55***	37.36***	6.22	−58.72***
	(18.10)	(12.57)	(63.66)	(13.12)	(7.93)	(20.35)
Number of observations	370	370	370	340	340	340
Number of Role B subjects	74	74	74	68	68	68

Note: Double-censored random effects Tobit regressions. The punishments are normalized as the percentage of the maximum available punishment in the respective treatment. Bootstrapped standard errors clustered by Role B participants (punishers) are in parentheses: $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table A2: Treatment Differences in Punishment: Controlling for Role B Characteristics

	Punishment in Monetary Equivalents			Punishment as Percentage of the Maximum Punishment		
	Take All	Take Half	Take Nothing	Take All	Take Half	Take Nothing
Non-monetary treatment (=1)	−1.36***	−0.64***	−0.22**	0.06	2.34	4.47
	(0.21)	(0.14)	(0.11)	(5.74)	(3.89)	(3.58)
Role B's (punisher) characteristics						
Slider Task Bonus	0.02	−0.07*	−0.05*	0.92	−1.07	−1.15
	(0.06)	(0.04)	(0.03)	(1.64)	(1.11)	(0.97)
Random Payment	−0.04	0.00	0.02	−1.40	−0.19	0.54
	(0.06)	(0.04)	(0.03)	(1.43)	(0.95)	(0.79)
Willingness to Pay	0.08	0.03	−0.04	2.61	1.44	0.19
	(0.08)	(0.05)	(0.04)	(1.86)	(1.25)	(1.15)
Donation decision (%)	0.04***	0.01**	0.00	0.99***	0.26**	0.00
	(0.01)	(0.00)	(0.00)	(0.18)	(0.13)	(0.11)
Age	−0.02	−0.00	0.01	0.12	0.22	0.17
	(0.02)	(0.01)	(0.01)	(0.60)	(0.39)	(0.33)
Female (=1)	−0.04	−0.19	−0.13	1.98	−1.78	0.18
	(0.24)	(0.16)	(0.12)	(6.54)	(4.37)	(3.59)
Cognitive Reflection Test (0–3)	−0.05	−0.09	−0.07*	−1.01	−2.10	−1.88
	(0.09)	(0.06)	(0.04)	(2.71)	(1.74)	(1.40)
Political views (0–5)	0.05	0.06	0.00	5.46*	3.92*	1.42
	(0.11)	(0.07)	(0.05)	(2.92)	(2.02)	(1.88)
Constant	1.92**	1.09**	0.44	4.97	4.41	3.67
	(0.78)	(0.45)	(0.37)	(21.47)	(13.55)	(11.76)
Number of observations	710	710	710	710	710	710
Number of Role B subjects	142	142	142	142	142	142

Note: Donation decision is the percentage of Total Earnings donated. Cognitive Reflection Test refers to the number of correct answers. Political views refers to the responses to question "What are your political views in general?": 1 Left; 2 Center Left; 3 Center; 4 Center Right; 5 Right. Robust standard errors clustered by Role B participants (punishers) are in parentheses: $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.