Human punishment is motivated by both a desire for revenge and a desire for equality

JONATHAN E BONE

CoMPLEX, University College London

NICHOLA J RAIHANI*

Department of Genetics, Evolution and Environment, University College London

*Corresponding author:

Nichola J Raihani

Department of Genetics, Evolution and Environment,

University College London,

London,

WC1E 6BT,

United Kingdom.

nicholaraihani@gmail.com

Abstract

Humans willingly pay a cost to punish defecting partners in experimental games. However, the psychological motives underpinning punishment are unclear. Punishment could stem from the desire to reciprocally harm a cheat (i.e. revenge) which is arguably indicative of a deterrent function. Alternatively, punishment could be motivated by the desire to redress the balance between punisher and cheat. Such a desire for equality might be more indicative of a fitness-leveling function**.** We used a two player experimental game to disentangle these two possibilities. In this game, one player could choose to steal $0.20 from their partner. Depending on the treatment, players interacting with a stealing partner experienced either advantageous inequality, equal outcomes or disadvantageous inequality. Players could punish stealing partners but some players had access to effective punishment (1 : 3 fee to fine) whereas others could only use ineffective punishment (1 : 1). Players who had access to effective punishment could reduce disadvantageous inequality by tailoring their investment in punishment whereas ineffective punishment did not change the relative payoffs of the individuals in the game but could be used to exact revenge. Players punished regardless of whether stealing created outcome inequality or whether punishment was ineffective at removing payoff differentials, suggesting that punishment was at least partly motivated by the desire to inflict reciprocal harm. However, in the effective punishment condition, players' tendency to punish increased if stealing resulted in disadvantageous inequality and, when possible, punishers tailored their investment in punishment to create equal outcomes. Together these findings suggest that punishment is motivated by both a desire for revenge and a desire for equality. The implications of these findings are discussed.

1. Introduction

Punishment typically involves paying a cost to harm individuals who harm or withhold benefits from the punisher (hereafter 'defectors', Clutton-Brock & Parker, 1995; Raihani, Thornton, & Bshary, 2012; but see Irwin & Horne, 2013; Sylwester, Herrmann, & Bryson, 2013 for punishment aimed at helpful or cooperative individuals). Since punishment is costly to administer, both in terms of executing the punishment itself and in terms of the possibility of provoking retaliation from the target (Dreber, Rand, Fudenberg, & Nowak, 2008; Herrmann, Thöni, & Gächter, 2008; Janssen & Bushman, 2008; Nikiforakis, 2008), considerable effort has been expended in trying to understand the evolved function of punitive sentiments (McCullough, Kurzban, & Tabak, 2013; Price, Cosmides, & Tooby, 2002). Specifically, it has been argued that understanding the contexts that reliably motivate punishment can provide key insights into its likely evolved function (Price et al., 2002). Two broad functional explanations have been proposed. First, it has been suggested that punitive sentiment could confer a selective advantage if punishment deters targets (or bystanders) from harming the punisher in future interactions (e.g. dos Santos, Rankin, & Wedekind, 2011; Hilbe & Sigmund, 2010; McCullough et al., 2013). Under this hypothesis (hereafter the 'revenge' hypothesis), individuals should be motivated to reciprocally harm individuals that intentionally harm them, even if punishment cannot immediately equalize the payoffs between the defector and the punisher (Falk, Fehr, & Fischbacher, 2005). However, evidence that punitive sentiments are sensitive to the risk of suffering a fitness disadvantage relative to defectors (Dawes, Fowler, Johnson, McElreath, & Smirnov, 2007; Raihani & McAuliffe, 2012a) suggests an alternative explanation: that punishment primarily serves a fitness-leveling function, by reducing payoff differentials between defectors and punishers (Price et al., 2002). Under this fitness-leveling hypothesis, punishers are expected to be motivated primarily by the desire to equalize payoffs and any deterrent function of punishment would arise as a by-product. Here, we present an experiment to test whether punitive sentiment can best be explained in terms of desire for revenge or in terms of a desire to equalize payoffs in social interactions.

Interacting with a defector often reduces cooperators' payoffs and creates unequal outcomes. It can therefore be difficult to establish whether punishment of defectors is motivated by the disutility

associated with receiving lower payoffs than a defector ('disadvantageous inequality aversion', (Fehr & Schmidt, 1999) or simply a desire for revenge (Raihani & McAuliffe, 2012b). A recent study attempted to disentangle these two possible motivations by asking whether, in the absence of disadvantageous inequality, experiencing losses was sufficient to motivate punishment (Raihani & McAuliffe, 2012b). Raihani & McAuliffe (2012b) found that defection, in the form of stealing money from the victim, did not motivate punishment when stealing resulted in equal outcomes or advantageous inequality for the victim. However, stealing did motivate punishment when it resulted in disadvantageous inequality for the victim (Raihani & McAuliffe, 2012b). These findings raise the possibility that individuals use punishment to restore equality in social interactions. However, the alternative possibility, that punishment is simply related to the disutility associated with experiencing disadvantageous inequality and is not tailored to achieve equal outcomes, could not be ruled out because players in this game were not allowed to tailor their investment in punishment.

Alternative studies have also suggested that investment in punishment is aimed at producing equal outcomes in social interactions. For example, in (Dawes et al., 2007) individuals were placed in groups of four and randomly allocated an endowment. Some players therefore started out richer than others in this game. Players were given the option to reduce (or increase) the income of others by purchasing negative (income-reducing) or positive (income-increasing) tokens and allocating these to other group members. In this setting, people allocated more negative tokens to the richest players and allocated more positive tokens to the poorest members of the group - suggesting that these behaviors were aimed at reducing outcome inequality. However, in this experiment, all four group members were able to purchase and allocate these tokens. Thus, it was impossible for players to predict how many tokens they would need to buy in order to achieve equal outcomes. Consequently, it is not possible to determine whether players adjusted investment in punitive behavior in order to achieve specific outcomes. Moreover, since initial payoff inequalities were exogenously determined rather than arising through some players defecting, the study could not test to what extent investment in income-reducing tokens was related to the partner's behavior, as opposed to the outcome itself. In other words, since

cooperation and defection were not possible in this game, any revenge-based motives of punishment could not be measured.

A more recent study by (Houser & Xiao, 2010) showed that players who were treated unfairly most commonly chose to punish as severely as possible and thus create inequality in their own favor. Although this seems to be more suggestive of punishment as a form of revenge rather than a fitness-leveler, it is important to take into account that in this study the severity of punishment chosen was not constrained by cost. In reality, imposing a larger cost on another individual is likely to also impose a larger cost on the punisher (Raihani & McAuliffe, 2012a). Since punishers have been shown to adjust their investment according to the costs associated with punishment (Anderson & Putterman, 2006; Bone, Silva, & Raihani, 2014; Carpenter, 2007; Nikiforakis & Normann, 2008; Ostrom, Walker, & Gardner, 1992), this creates a potentially important trade-off between maximizing income and achieving the desired punishment outcome.

The fitness-leveling hypothesis predicts that individuals should only invest in punishment that is more costly to the target than to the punisher, and is therefore able to reduce any existing disadvantageous inequality. Nevertheless, empirical work has demonstrated that individuals are prepared to invest in punishment that is equally costly to the punisher and the target (Anderson & Putterman, 2006; Carpenter, 2007; Egas & Riedl, 2008; Falk et al., 2005; Nikiforakis & Normann, 2008) - or even more costly to the punisher (Anderson & Putterman, 2006; Carpenter, 2007; Egas & Riedl, 2008) - and so is unable to re-establish equality. These findings suggest that punishers are not solely motived by a desire to remove fitness differentials and support the idea that punishers might instead be motivated by a desire for revenge against defecting partners. The predictions of the two hypotheses also differ with respect to whether the defection was performed intentionally or not. Specifically, the revenge hypothesis predicts that punishment should be focused on those who impose harm intentionally and can therefore learn to avoid repeating the harmful behavior in the future. Conversely, punishment aimed at removing fitness differentials should be less sensitive (or insensitive) to intentionality since the primary function is to reduce inequality rather than change the target's behavior. Evidence from empirical studies provides

some support for both hypotheses. Whilst several studies have shown that individuals will punish in response to unequal outcomes created at random or unintentionally (Cushman, Dreber, Wang, & Costa, 2009; Dawes et al., 2007; Falk, Fehr, & Fischbacher, 2008; Houser & Xiao, 2010; Kagel, Kim, & Moser, 1996; Yu, Calder, & Mobbs, 2014), individuals are significantly more likely to punish when unequal outcomes are created intentionally by the target (Falk et al., 2008; Houser & Xiao, 2010; Kagel et al., 1996).

Based on past research it is therefore unclear whether punishment is motivated by a desire for revenge or by a desire to equalize payoffs. We aimed to answer this question by investigating whether victims of cheats adjusted their investment in punishment in order to restore equality using a modified version of the game used by (Raihani & McAuliffe, 2012b). In this game, one player could choose to steal $0.20 from their partner. Depending on the treatment, players interacting with a stealing partner experienced advantageous inequality, equal outcomes or varying levels of disadvantageous inequality. Players could punish stealing partners but while some players had access to effective punishment (1 : 3 fee to fine) others could only use ineffective punishment (1 : 1 fee to fine). Players who had access to effective punishment could achieve equal outcomes by tailoring their investment in punishment: more extreme outcome inequality could be alleviated by investing more into punishment. However, under the ineffective punishment condition, increasing investment in punishment did not reduce inequality.

Although we suggest that revenge may serve a deterrent function, in the anonymous one-shot setting of our game, there is no scope for punishment to change the behavior of stealing partners (or bystanders). However, previous work has suggested that behavior may be constrained by psychological mechanisms that evolved in the context of non-anonymous repeated interactions and that responses that are attuned to these conditions may be invoked even in anonymous, one-shot settings (Ben-Ner & Putterman, 2000; Burnham & Johnson, 2005; Cosmides & Tooby, 1989; Delton, Krasnow, Cosmides, & Tooby, 2011; Hagen & Hammerstein, 2006; Hoffman, McCabe, & Smith, 1998; Johnson, Stopka, & Knights, 2003; Tooby, Cosmides, & Price, 2006). Thus, in our game a desire for revenge might reflect the desires of an evolved psychology that functions to deter cheats, even though this function is (due to the

nature of the game) impossible to achieve. Nevertheless, we note that since deterrence is not the only possible function for this behavior we use the word 'revenge' in a purely descriptive sense.

The revenge hypothesis predicts that punishment will be used in both the ineffective and the effective punishment condition. Alternatively, if punishment is motivated by the desire to equalize outcomes, punishment should be used when it is effective but not when it is ineffective. Moreover, players should use the amount of punishment that is required to equalize payoffs (Table 1); not more or less.

2. Materials and Methods

2.1 Experimental protocol

This research was approved by the University College London ethics board project number 3720/001. Data were collected in October 2013 and July - August 2014. We recruited 4912 subjects (2856 males, 1967 females, 89 unreported) for our experiment using the online labor market, Amazon Mechanical Turk (AMT; www.mturk.com). Subjects were all based in the USA.

Of the 4912 subjects, 2456 were assigned the role of player one (P1). The remaining subjects were allocated the role of player two (P2). P1 and P2 were both allocated one of five initial endowments (treatment A – E; Table 1). The game consisted of two stages. In the first stage P2 could choose to steal \$0.20 from P1 or do nothing. In the second stage, P1 was informed of P2's decision and could choose how many punishment points they wished to assign to P2. P1 experienced the same losses when P2 stole (\$0.20) in all five treatments. However, depending on the treatment this \$0.20 loss resulted in P1 experiencing either advantageous inequality (treatment A), equal payoffs (treatment B) or disadvantageous inequality (treatments C – E; Table 1) relative to P2. All players were assigned to one of two punishment conditions at the start of the game: ineffective and effective. In the ineffective punishment condition, each punishment point cost P1 \$0.05 and reduced P2's earnings by \$0.05 (fee to fine ratio = 1 : 1). In the effective punishment condition, each punishment point cost P1 \$0.05 and reduced P2's earnings by \$0.15 (fee to fine ratio = 1 : 3). To prevent negative earnings, P1 could assign a maximum of four punishment points to P2.

P1 was assigned ex-post (Rand, 2012) to one of two treatments in which either P2 stole or P2 didn't steal (Table 1). These treatments were allocated to players both in the ineffective and effective punishment conditions, creating a total of 10 treatments for P2 and a total of 20 treatments for P1. All subjects that participated in the experiment received a $0.20 show-up payment on top of a bonus based on both their and their partner's decisions during the game.

2.2 Analysis

Data were analyzed using R version 2.15.2 (R Development Core Team, 2011). All comparisons used two-sided Fishers exact tests. Firstly, we investigated whether experiencing losses or disadvantageous inequality had a greater effect on P1's decision to punish P2. We compared the proportion of P1s that chose a non-zero punishment investment when (i) P2 didn't steal (across all treatments), (ii) P2 stole but the stealing did not result in disadvantageous inequality for P1 (i.e. treatments A & B) and (iii) P2 stole resulting in disadvantageous inequality for P1 (i.e. treatments C – E). Separate analyses were conducted for players in the effective and players in the ineffective punishment conditions (see Table 2 for comparisons and sample sizes).

Next, we investigated whether the inequality-removing punishment investment was picked more frequently than each of the other three possible punishment investments. Data were restricted to instances where P1s punished P2 for stealing in treatments where P2 stealing created disadvantageous inequality for P1 (treatments C – E) and when P1 had access to effective punishment (see Table 3 for comparisons and sample sizes). We then asked whether these punitive players were less willing to invest the amount required to create equality when doing so became progressively more expensive (table 1).

Finally, we investigated the possibility that players that chose the inequality-removing punishment investment may have done so because that amount of punishment was related to the disutility associated with the level of inequality experienced in that treatment, even when punishment was incapable of restoring equality (i.e. when punishment was ineffective). For this analysis we compared the proportion of

players in the effective punishment condition that chose the inequality-removing punishment investment versus the proportion of players in the ineffective punishment condition that chose that same punishment investment. Data were restricted to instances where P1s punished P2 for stealing in treatments where P2 stealing created disadvantageous inequality for P1 (treatments C – E).

As multiple comparisons were performed, sequential Benjamini - Hochberg adjusted p^{BH}-values (Benjamini & Hochberg, 1995; see also Waite & Campbell, 2006) are reported alongside uncorrected p-values. By controlling for the false discovery rate, Benjamini - Hochberg adjusted p-values balance the risk of incurring Type I errors with the risk of incurring Type II errors.

3. Results

In both the effective and the ineffective punishment condition, P1 was significantly more likely to punish a stealing than a non-stealing P2 (Fisher's exact test, see Table 2 for p-values; Figure 1). In the effective punishment condition, the tendency to punish a stealing P2 was increased significantly when stealing resulted in disadvantageous inequality (proportion punishing non-stealing P2 ± SE = 0.04 ± 0.01; stealing P2, no disadvantageous inequality = 0.19 ± 0.03; stealing P2, disadvantageous inequality = 0.34 ± 0.02; Table 2; Figure 1). Although players in the ineffective punishment condition also appeared to be more likely to punish a stealing P2 when stealing resulted in disadvantageous inequality, this finding was non-significant (proportion punishing non-stealing P2 ± SE = 0.03 ± 0.01; stealing P2, no disadvantageous inequality = 0.06 ± 0.01; stealing P2, disadvantageous inequality = 0.1 ± 0.02; Table 2; Figure 1).

When P2 stealing created disadvantageous inequality for P1 (treatments C - E), if P1 had access to effective punishment, P1 could equalize outcomes by punishing P2. The specific punishment investment that would create equal outcomes depended on the treatment (Table 1). In treatments C - E, when punishment was effective, the punishment investment that created equal outcomes was chosen significantly more often than any other possible investment (Fisher's exact test, see Table 3 for p-values; Figure 2); and this punishment investment was chosen significantly more frequently in the effective than

the ineffective punishment condition (Fisher's exact test, p-value = 0.024; p^{BH}-value = 0.027; n = 167; Figure 2). Moreover, in the effective punishment condition, a punishing P1 was equally likely to choose the punishment investment that created equal outcomes in all three treatments where P2 stealing created disadvantageous inequality for P1 (C – E; Fisher's exact test, p-value **=** 0.698; p^{BH}-value = 0.698; n = 130; Figure 2), indicating that players' attempts to equalize outcomes were largely insensitive to the cost associated with doing so. All significant findings reported above remained significant after p-values were adjusted according to the Benjamini–Hochberg procedure (Benjamini & Hochberg, 1995).

When punishment was effective, in both treatments where P2 stealing did not create disadvantageous inequality for P1 (treatments A & B), if P1 used punishment, they were most likely to choose the harshest punishment option available (Figure 2; see ESM for details). Due to the small proportion of P1 who chose to punish P2 when P2 didn't steal or when punishment was ineffective, we did not have the statistical power to test which punishment investments were most popular in these scenarios (see ESM for power analysis, conducted using GPower; Erdfelder, Faul, & Buchner, 1996).

4. Discussion

In this study, P1 experienced the same losses when P2 stole ($0.20) in all five treatments. However, depending on the treatment this $0.20 loss resulted in P1 experiencing either advantageous inequality (treatment A), equal payoffs (treatment B) or disadvantageous inequality (treatment C - E) relative to P2. P2 stealing provoked P1 to punish even when stealing did not create disadvantageous inequality. Moreover, although relatively rare, P1 sometimes punished a stealing P2 even when punishment was ineffective and was thus unable to re-establish equality. Both these findings suggest that punishment was motivated at least in part by a desire for revenge against stealing partners. However, when punishment was effective, P1 was more likely to punish if P2 stealing created disadvantageous inequality and, when given the option, P1 typically adjusted their investment in punishment to create equal outcomes. This suggests that although a desire for revenge was sometimes sufficient to motivate punishment, players were also sensitive to inequality and preferred punishment to result in equal outcomes.

Previous studies using three-player games have also shown that players will use apparently ineffective punishment (Egas & Riedl, 2008; Falk et al., 2005). However, in these studies, it could be argued that, although ineffective punishment does not reduce inequality between players, it can reduce the standard deviation of the group's mean payoff and so may still be driven by egalitarian motives (Dawes et al., 2007). This, however, is not possible in two-player games like ours. In the current study, a willingness to pay for ineffective punishment therefore seems to reflect a desire for revenge (with an associated deterrent function, (McCullough et al., 2013) despite the fact that punishment occurred in an anonymous, one-shot setting where no deterrent function was possible. Although in our game punishment yielded no potential return on investment for punishers in terms of changing the partner's future behavior), previous studies have proposed that the psychological mechanisms that underpin social behavior (e.g. punishment) are likely to have evolved in a context where one-shot or anonymous interactions were rare (Delton et al., 2011; Fehr & Henrich, 2003). It has been suggested that this evolved psychology may invoke responses that are attuned to these conditions even in that are not adaptive in truly anonymous, one-shot lab settings (Ben-Ner & Putterman, 2000; Burnham & Johnson, 2005; Cosmides & Tooby, 1989; Delton et al., 2011; Fehr & Henrich, 2003; Hagen & Hammerstein, 2006; Hoffman et al., 1998; Johnson et al., 2003; Tooby et al., 2006). Thus, it is possible that the use of ineffective punishment in our game was caused by the mis-firing of psychological mechanisms adapted to deter defecting partners from future defection, even when this function is (due to the nature of the game) impossible to achieve. Previous work has shown that when players are put under time pressure to make decisions in one-shot games, they are more likely behave cooperatively (Rand et al., 2014; Rand, Greene, & Nowak, 2012). Similarly, other studies also using one-shot games have shown that when players are given a cooling off period they are less likely to punish cheating partners (Grimm & Mengel, 2011; Smith & Silberberg, 2010; Sutter, Kocher, & Strauß, 2003). These studies suggest that when players are the given time to consider their decisions they are more likely to respond in a way that maximizes their payoff in their current one-shot setting rather than rely on intuitions that may maximize payoffs over repeated encounters in the real world but not one-shot laboratory settings.

On the other hand, it could also be argued that the proximate mechanisms that underpin punishment may have evolved in a context where punishment was likely to have imposed larger costs on the target than the punisher (i.e. punishment was effective) and was therefore capable of reducing the disadvantageous inequality experienced by victims of cheats. This line of reasoning would lead to the conclusion that a willingness to invest in ineffective punishment in our game could reflect the desires of an evolved psychology with the function of leveling fitness differentials (even though this function cannot be achieved in a context where punishment is ineffective). It is currently not clear what the most realistic fee to fine ratio is to use for punishment in laboratory settings in order to approximate the cost to impact ratio of punishment under real-world settings. Indeed, under real-world settings, the fee-to-fine ratio of punishment is likely to vary with relative dominance status of individuals (e.g. Bone, Wallace, Bshary, & Raihani, 2015; Raihani et al., 2012). Clearly, more studies of punishment in real-world settings are needed to establish how punishment use varies according to whether interactions are repeated or not; and whether the fee to fine ratios currently used in the laboratory studies are ecologically valid.

Several players in this study used punishment to create advantageous inequality in their favor. For example, when players had access to effective punishment and faced a stealing partner without also experiencing disadvantageous inequality (treatments A & B), punishing P1s typically chose the punishment investment that created the largest advantageous inequality for themselves. This finding is consistent with the idea that punishment is motivated by a desire for revenge, which might be 'sweeter' the more it harms the target (de Quervain et al., 2004); and is comparable to the findings from Houser & Xiao's (2010) study which showed that when the severity of punishment used was not constrained by cost, players chose to punish as severely as possible and thus create inequality in their own favor (Abbink & Sadrieh, 2009). This finding supports the idea that punishment is motivated by competitive motives, where players value being in a position of advantageous inequality because it emphasizes their relative social status (Fershtman, Gneezy, & List, 2012; Houser & Xiao, 2010).

Similar competitive motives have been inferred for the existence of 'antisocial' punishment (Herrmann et al., 2008; Raihani & Bshary, 2015; Sylwester et al., 2013). As in several previous studies

(e.g. Anderson & Putterman, 2006; Gächter, Herrmann, & Thöni, 2005; Gächter & Herrmann, 2009; Herrmann et al., 2008), we documented antisocial punishment (aimed at non-stealing partners) in this study. In this context, antisocial punishment cannot be explained by a desire for revenge or a desire to reduce disadvantageous inequality since P1 experienced neither losses nor disadvantageous inequality when P2 did not steal. It may be the case that antisocial punishment reflects competitive motives (Prediger, Vollan, & Herrmann, 2014; Raihani & Bshary, 2015; Sylwester et al., 2013), though if this were the case we would have expected that P1s would use have used antisocial punishment in the effective but not in the ineffective punishment condition, as previously documented (Falk et al., 2005). In contrast to this prediction, we found that players were equally likely to punish antisocially regardless of the punishment condition. It is possible that antisocial punishment in this study simply reflects execution errors or misperceiving the game. With the current dataset we are unable to determine the causes of antisocial punishment but this remains an exciting avenue for future research.

Although many of our results support the idea that punishment was motivated primarily by a desire for revenge, we report two findings that support the hypothesis that punishment is motivated by a desire for equality (with an associated fitness-leveling function (Price et al., 2002). First, as in Raihani & McAuliffe (2012b), we found that in the effective punishment condition players' tendency to punish a stealing partner was increased if stealing resulted in resulted in disadvantageous inequality. Second, when given the option, players typically tailored their investment in effective punishment to remove disadvantageous inequality and were seemingly insensitive to the cost associated with achieving this outcome. Moreover, the punishment investment that created equal outcomes was chosen much more frequently in the effective than the ineffective punishment condition, indicating that players were attempting to create equal outcomes rather than increasing punishment investment in response to frustration at experiencing increasingly disadvantageous outcomes.

Together our findings suggest that punishment is motivated by both a desire for revenge and a desire for equality. Indeed, these possibilities are not mutually exclusive. Furthermore, it might be the case that punishment which results in equality may be most likely to serve a deterrent function, if such

punishment is perceived to be ‘fair’ and consequently more effective at changing the target's behavior. This prediction is based on previous studies, where colleagues have suggested that 'morally legitimate' punishment is most likely to successfully deter future defection (Fehr & Rockenbach, 2003; Houser & Xiao, 2010). Fehr & Rockenbach (2003) suggest that punishment may be perceived as being morally illegitimate if it is associated with selfish or greedy (rather than altruistic) intentions. Punishment that creates advantageous inequality in the punisher's favor might be interpreted as a competitive act (Raihani & Bshary, 2015) and therefore perceived as morally illegitimate. Punishment that creates advantageous inequality in favor of the punisher might therefore be unlikely to deter further defection (Bone et al., 2015; Fehr & Rockenbach, 2003; Xiao, 2013) and may even provoke retaliation from the target (Bone et al., 2015). Whilst we stress that this explanation is speculative it offers promising avenues for further studies to explore the scenarios that motivate punishment.

We note that the current findings appear to contradict the results of Raihani & McAuliffe (2012b), who showed, using a similar experimental setup, that P1 only punished P2 where P2 stealing resulted in disadvantageous inequality. In contrast, in the current study, we found that players punished stealing partners even when stealing did not create disadvantageous inequality. We believe it is unlikely that the different costs of punishment used in the two studies ($0.05 in this study; $0.10 in the previous study) are responsible for these conflicting results (see ESM for supporting analysis). However, it is possible that other subtle differences between our experimental setups may be responsible, specifically differences in the endowments initially given to P1 or differences in the demographic sample across the studies. In Raihani & McAuliffe’s (2012b) experiment, the losses experienced by P1 as a result of P2 stealing ($0.20) were the same as in this experiment. However, the initial endowment of P1 was different: in Raihani & McAuliffe (2012b), P1 began the game with $0.70 and was left with $0.50 if P2 stole, whereas in this study, P1 began the game with $1.10 and was left with $0.90 if P2 stole. It has been shown that people pay most attention to the left-most digits when judging differences in the magnitude of numbers; a phenomenon known as the left-digit anchoring effect (Dehaene, Dupoux, & Mehler, 1990; Hinrichs, Yurko, & Hu, 1981; Monroe & Lee, 1999; Thomas & Morwitz, 2005). For example, an experimental study showed that a reduction of one cent affected the perceived magnitude of a price when the left digit changed

($3.00 to $2.99) but not when the left digit was unchanged ($3.20 to $3.19) (Thomas & Morwitz, 2005). Thus, a reduction from $1.10 to $0.90 (as P1 experienced in this study) may be perceived as a greater loss than a reduction from $0.70 to $0.50 (as P1 experienced in Raihani & McAuliffe (2012b). If players perceived greater losses in this study than in Raihani & McAuliffe (2012b), this may explain why a $0.20 loss which did not result in disadvantageous inequality motivated P1 to punish P2 in this study but not in the earlier Raihani & McAuliffe (2012b) experiment. In other words, in the absence of unequal outcomes, the loss experienced by P1 as a result of P2 stealing in Raihani & McAuliffe (2012b) may have been perceived as too small to motivate punishment.

Alternatively, the discrepancy between the results of the current study and Raihani & McAuliffe (2012b) may be explained by differences in demographic sampling between the two studies. Data for both studies were collected via the online labor market, Amazon Mechanical Turk, where the vast majority of workers hail from either the USA or India (Ross, Irani, Silberman, Zaldivar, & Tomlinson, 2010). In the Raihani & McAuliffe (2012b) study, participants were recruited from both countries and the analysis did not control for the possible cross-cultural differences in subjects' behavior. However, in this study we restricted participation to subjects based in the USA. Previous studies have demonstrated cross-cultural differences in the propensity of subjects to punish both defectors (Henrich et al., 2006; Marlowe & Berbesque, 2008) and cooperative individuals (Ellingsen, Herrmann, Nowak, Rand, & Tarnita, 2012; Gächter & Herrmann, 2009; Herrmann et al., 2008). Thus, differences in the way that subjects from India versus the US behave in economic games, particularly with respect to punishment, may explain the different results we saw across the two studies. Future work will explore how cultural differences between players affect punishment strategies.

To summarize, we investigated whether punishment was motivated by a desire for revenge or a desire for equality and found support for both of these hypotheses. Players used punishment regardless of whether stealing created outcome inequality or whether punishment was ineffective at removing payoff differentials. This supports the hypothesis that punishment is motivated by revenge. However, players were more likely to punish if stealing resulted in disadvantageous inequality for the punisher and, when

possible, typically tailored their investment in punishment to create equal outcomes. This supports the hypothesis that punishment is motivated by a desire to equalize payoffs. Since these hypotheses are not mutually exclusive we suggest that both a desire for revenge and a desire for equality are likely to play an important role in motivating punishment decisions. Future work should explore how the efficacy of punishment is related to its perceived moral legitimacy, and whether players are sensitive to this when tailoring investment in punishment. We also suggest that more work is needed to understand what motivates antisocial punishment, how intuitions guide punishment decisions and in what ways cultural variation between players influences punishment strategies.

5. Acknowledgements

We would like to thank the editor, Robert Kurzban, and the three anonymous referees for their useful comments. This study was funded by a Royal Society University Research Fellowship to N.R.

6. References

Abbink, K., & Sadrieh, A. (2009). The pleasure of being nasty. *Economics Letters*, *105*(3), 306–308. doi:10.1016/j.econlet.2009.08.024

Anderson, C. M., & Putterman, L. (2006). Do non-strategic sanctions obey the law of demand? The demand for punishment in the voluntary contribution mechanism. *Games and Economic Behavior*, *54*(1), 1–24. doi:10.1016/j.geb.2004.08.007

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, *57*(1), 289 – 300. doi:10.2307/2346101

Ben-Ner, A., & Putterman, L. (2000). On some implications of evolutionary psychology for the study of preferences and institutions. *Journal of Economic Behavior & Organization*, *43*(1), 91–99.

Bone, J., Silva, A. S., & Raihani, N. J. (2014). Defectors, not norm violators, are punished by third-parties. *Biology Letters*, *10*(7), 20140388–. doi:10.1098/rsbl.2014.0388

Bone, J., Wallace, B., Bshary, R., & Raihani, N. J. (2015). The Effect of Power Asymmetries on Cooperation and Punishment in a Prisoner's Dilemma Game. *PloS One*, *10*(1), e0117183. doi:10.1371/journal.pone.0117183

Burnham, T. C., & Johnson, D. D. P. (2005). The biological and evolutionary logic of human cooperation.

Carpenter, J. P. (2007). The demand for punishment. *Journal of Economic Behavior & Organization*, *62*(4), 522–542. doi:10.1016/j.jebo.2005.05.004

Clutton-Brock, T. H., & Parker, G. A. (1995). Punishment in animal societies. *Nature*, *373*(6511), 209–216. doi:10.1038/373209a0

Cosmides, L., & Tooby, J. (1989). Evolutionary psychology and the generation of culture, part II: Case study: A computational theory of social exchange. *Ethology and Sociobiology*, *10*(1–3), 51–97.

Cushman, F., Dreber, A., Wang, Y., & Costa, J. (2009). Accidental outcomes guide punishment in a “trembling hand” game. *PloS One*, *4*(8), e6699. doi:10.1371/journal.pone.0006699

Dawes, C. T., Fowler, J. H., Johnson, T., McElreath, R., & Smirnov, O. (2007). Egalitarian motives in humans. *Nature*, *446*(7137), 794–6. doi:10.1038/nature05651

De Quervain, D. J.-F., Fischbacher, U., Treyer, V., Schellhammer, M., Schnyder, U., Buck, A., & Fehr, E. (2004). The neural basis of altruistic punishment. *Science (New York, N.Y.)*, *305*(5688), 1254–8. doi:10.1126/science.1100735

Dehaene, S., Dupoux, E., & Mehler, J. (1990). Is numerical comparison digital? Analogical and symbolic effects in two-digit number comparison. *Journal of Experimental Psychology. Human Perception and Performance*, *16*(3), 626–41.

Delton, A. W., Krasnow, M. M., Cosmides, L., & Tooby, J. (2011). Evolution of direct reciprocity under uncertainty can explain human generosity in one-shot encounters. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(32), 13335–40. doi:10.1073/pnas.1102131108

Dos Santos, M., Rankin, D. J., & Wedekind, C. (2011). The evolution of punishment through reputation. *Proceedings. Biological Sciences / The Royal Society*, *278*(1704), 371–377.

Dreber, A., Rand, D. G., Fudenberg, D., & Nowak, M. A. (2008). Winners don’t punish. *Nature*, *452*(7185), 348–51. doi:10.1038/nature06723

Egas, M., & Riedl, A. (2008). The economics of altruistic punishment and the maintenance of cooperation. *Proceedings. Biological Sciences / The Royal Society*, *275*(1637), 871–8. doi:10.1098/rspb.2007.1558

Ellingsen, T., Herrmann, B., Nowak, M. A., Rand, D. G., & Tarnita, C. E. (2012). Civic Capital in Two Cultures: The Nature of Cooperation in Romania and USA. *CESifo Working Paper Series*.

Erdfelder, E., Faul, F., & Buchner, A. (1996). GPOWER: A general power analysis program. *Behavior Research Methods, Instruments, & Computers*. doi:10.3758/BF03203630

Falk, A., Fehr, E., & Fischbacher, U. (2005). Driving Forces Behind Informal Sanctions. *Econometrica*, *73*(6), 2017–2030. doi:10.1111/j.1468-0262.2005.00644.x

Falk, A., Fehr, E., & Fischbacher, U. (2008). Testing theories of fairness--Intentions matter. *Games and Economic Behavior*, *62*(1), 287–303.

Fehr, E., & Henrich, J. (2003). Is Strong Reciprocity a Maladaptation? On the Evolutionary Foundations of Human Altruism. *SSRN Electronic Journal*. doi:10.2139/ssrn.382950

Fehr, E., & Rockenbach, B. (2003). Detrimental effects of sanctions on human altruism. *Nature*, *422*(6928), 137–40. doi:10.1038/nature01474

Fehr, E., & Schmidt, K. M. (1999). A Theory of Fairness, Competition, and Cooperation. *The Quarterly Journal of Economics*, *114*(3), 817–868. doi:10.1162/003355399556151

Fershtman, C., Gneezy, U., & List, J. A. (2012). Equity Aversion: Social Norms and the Desire to Be Ahead. *American Economic Journal: Microeconomics*, *4*(4), 131–44.

Gächter, S., & Herrmann, B. (2009). Reciprocity, culture and human cooperation: previous insights and a new cross-cultural experiment. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *364*(1518), 791–806. doi:10.1098/rstb.2008.0275

Gächter, S., Herrmann, B., & Thöni, C. (2005). Cross-cultural differences in norm enforcement. *Behavioral and Brain Sciences*, *28*(06), 822–823. doi:10.1017/S0140525X05290143

Grimm, V., & Mengel, F. (2011). Let me sleep on it: Delay reduces rejection rates in ultimatum games. *Economics Letters*, *111*(2), 113–115.

Hagen, E. H., & Hammerstein, P. (2006). Game theory and human evolution: a critique of some recent interpretations of experimental games. *Theoretical Population Biology*, *69*(3), 339–48. doi:10.1016/j.tpb.2005.09.005

Henrich, J., McElreath, R., Barr, A., Ensminger, J., Barrett, C., Bolyanatz, A., … Ziker, J. (2006). Costly punishment across human societies. *Science (New York, N.Y.)*, *312*(5781), 1767–70. doi:10.1126/science.1127333

Herrmann, B., Thöni, C., & Gächter, S. (2008). Antisocial punishment across societies. *Science (New York, N.Y.)*, *319*(5868), 1362–7. doi:10.1126/science.1153808

Hilbe, C., & Sigmund, K. (2010). Incentives and opportunism: from the carrot to the stick. *Proceedings. Biological Sciences / The Royal Society*, *277*(1693), 2427–2433.

Hinrichs, J. V., Yurko, D. S., & Hu, J. (1981). Two-digit number comparison: Use of place information. *Journal of Experimental Psychology: Human Perception and Performance*. doi:10.1037/0096-1523.7.4.890

Hoffman, E., McCabe, K. a., & Smith, V. L. (1998). Behavioral Foundations of Reciprocity: Experimental Economics and Evolutionary Psychology. *Economic Inquiry*, *36*(3), 335–352.

Houser, D., & Xiao, E. (2010). *Inequality-seeking punishment*. *Economics Letters* (Vol. 109, pp. 20–23).

Irwin, K., & Horne, C. (2013). A normative explanation of antisocial punishment. *Social Science Research*, *42*(2), 562–570.

Janssen, M. A., & Bushman, C. (2008). Evolution of cooperation and altruistic punishment when retaliation is possible. *Journal of Theoretical Biology*, *254*(3), 541–5. doi:10.1016/j.jtbi.2008.06.017

Johnson, D. D. P., Stopka, P., & Knights, S. (2003). Sociology: The puzzle of human cooperation. *Nature*, *421*(6926), 911–2; discussion 912. doi:10.1038/421911b

Kagel, J. H., Kim, C., & Moser, D. (1996). Fairness in ultimatum games with asymmetric information and asymmetric payoffs. *Games and Economic Behavior*, *13*(1), 100–110. doi:10.1006/game.1996.0026

Marlowe, F. W., & Berbesque, J. C. (2008). More “altruistic” punishment in larger societies. *Proceedings. Biological Sciences / The Royal Society*, *275*(1634), 587–590.

McCullough, M. E., Kurzban, R., & Tabak, B. a. (2013). Cognitive systems for revenge and forgiveness. *The Behavioral and Brain Sciences*, *36*(1), 1–15. doi:10.1017/S0140525X11002160

Monroe, K. B., & Lee, A. Y. (1999). Remembering versus Knowing: Issues in Buyers’ Processing of Price Information. *Journal of the Academy of Marketing Science*, *27*(2), 207–225. doi:10.1177/0092070399272006

Nikiforakis, N. (2008). Punishment and counter-punishment in public good games: Can we really govern ourselves. *Journal of Public Economics*, 91 – 112.

Nikiforakis, N., & Normann, H.-T. (2008). A comparative statics analysis of punishment in public-good experiments. *Experimental Economics*, *11*(4), 358–369. doi:10.1007/s10683-007-9171-3

Ostrom, E., Walker, J., & Gardner, R. (1992). Covenants With and Without a Sword: Self-Governance is Possible. *American Political Science Review*, *86*(2), 404–417. doi:10.2307/1964229

Prediger, S., Vollan, B., & Herrmann, B. (2014). Resource scarcity and antisocial behavior. *Journal of Public Economics*, *119*(119), 1–9. doi:10.1016/j.jpubeco.2014.07.007

Price, M. E., Cosmides, L., & Tooby, J. (2002). Punitive sentiment as an anti-free rider psychological device. *Evolution and Human Behavior*, *23*(3), 203–231.

R Development Core Team, R. (2011). R: A Language and Environment for Statistical Computing. (R. D. C. Team, Ed.)*R Foundation for Statistical Computing*. R Foundation for Statistical Computing. doi:10.1007/978-3-540-74686-7

Raihani, N. J., & Bshary, R. (2015). The reputation of punishers. *Trends in Ecology & Evolution*. doi:10.1016/j.tree.2014.12.003

Raihani, N. J., & McAuliffe, K. (2012a). Does Inequity Aversion Motivate Punishment? Cleaner Fish as a Model System. *Social Justice Research*, *25*(2), 213–231. doi:10.1007/s11211-012-0157-8

Raihani, N. J., & McAuliffe, K. (2012b). Human punishment is motivated by inequity aversion, not a desire for reciprocity. *Biology Letters*, (July), 18–21. doi:10.1098/rsbl.2012.0470

Raihani, N. J., Thornton, A., & Bshary, R. (2012). Punishment and cooperation in nature. *Trends in Ecology & Evolution*, *27*(5), 288–95. doi:10.1016/j.tree.2011.12.004

Rand, D. G. (2012). The promise of Mechanical Turk: how online labor markets can help theorists run behavioral experiments. *Journal of Theoretical Biology*, *299*(null), 172–9. doi:10.1016/j.jtbi.2011.03.004

Rand, D. G., Greene, J. D., & Nowak, M. A. (2012). Spontaneous giving and calculated greed. *Nature*, *489*(7416), 427–30. doi:10.1038/nature11467

Rand, D. G., Peysakhovich, A., Kraft-Todd, G. T., Newman, G. E., Wurzbacher, O., Nowak, M. A., & Greene, J. D. (2014). Social heuristics shape intuitive cooperation. *Nature Communications*, *5*, 3677. doi:10.1038/ncomms4677

Ross, J., Irani, I., Silberman, M. S., Zaldivar, A., & Tomlinson, B. (2010). Who are the Crowdworkers?: Shifting Demographics in Amazon Mechanical Turk. *Proc Extended Abstracts of the SIGCHI Conference on Human Factors in Computing Systems CHI*, 2863–2872.

Smith, P., & Silberberg, A. (2010). Rational maximizing by humans (Homo sapiens) in an ultimatum game. *Animal Cognition*, *13*(4), 671–7. doi:10.1007/s10071-010-0310-4

Sutter, M., Kocher, M., & Strauß, S. (2003). Bargaining under time pressure in an experimental ultimatum game. *Economics Letters*, *81*(3), 341–347. doi:10.1016/S0165-1765(03)00215-5

Sylwester, K., Herrmann, B., & Bryson, J. (2013). Homo homini lupus? Explaining antisocial punishment. *British Educational Research Journal*, *6*, 167–185. doi:<A id=link10.1037npe0000009 title="Search for 10.1037/npe0000009" href="javascript:__doLinkPostBack('','ss%7E%7EDI%20%2210%2E1037%2Fnpe0000009%22%7C%7Csl%7E%7Erl','');">10.1037/npe0000009</A>

Thomas, M., & Morwitz, V. (2005). Penny Wise and Pound Foolish: The Left-Digit Effect in Price Cognition. *Journal of Consumer Research*, *32*(1), 54–64. doi:10.1086/429600

Tooby, J., Cosmides, L., & Price, M. E. (2006). Cognitive Adaptations for n-person Exchange: The Evolutionary Roots of Organizational Behavior. *Managerial and Decision Economics : MDE*, *27*(2-3), 103–129. doi:10.1002/mde.1287

Waite, T. A., & Campbell, L. G. (2006). Controlling the false discovery rate and increasing statistical power in ecological studies. *Ecoscience*, *13*(4), 439–442. doi:10.2980/1195-6860(2006)13[439:CTFDRA]2.0.CO;2

Xiao, E. (2013). Profit-seeking punishment corrupts norm obedience. *Games and Economic Behavior*, *77*(1), 321–344. doi:10.1016/j.geb.2012.10.010

Yu, R., Calder, A. J., & Mobbs, D. (2014). Overlapping and distinct representations of advantageous and disadvantageous inequality. *Human Brain Mapping*, *35*(7), 3290–3301. doi:10.1002/hbm.22402

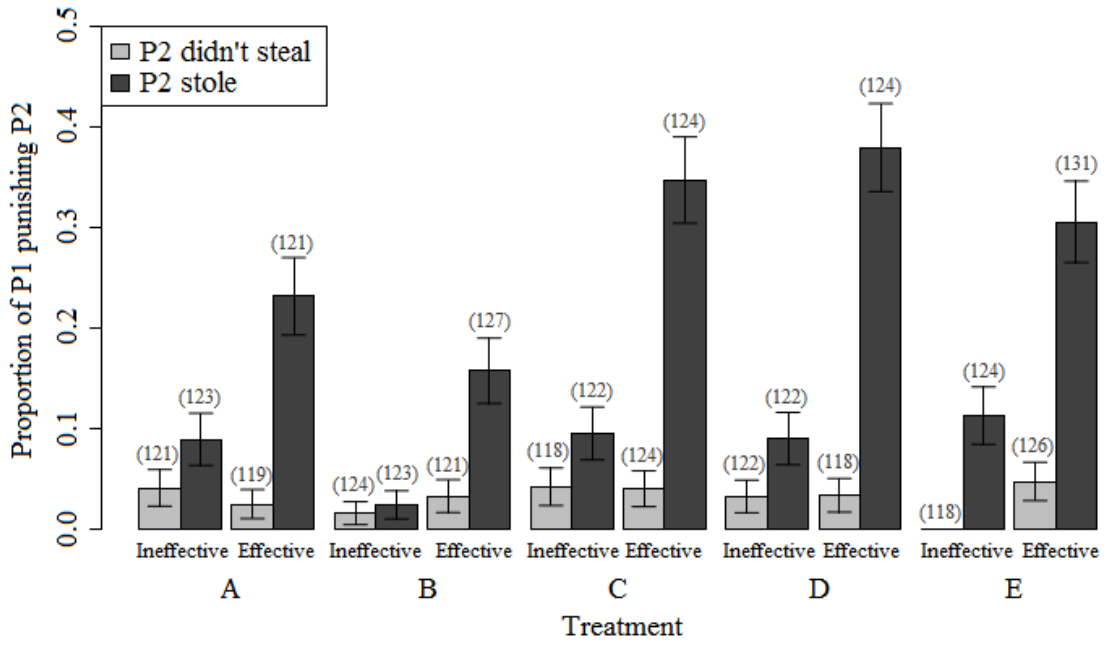

Figure 1. The proportion of P1 who punished P2 according to whether P2 stole (by taking $0.20 of P1's endowment), whether punishment was effective (fee to fine ratio = 1 : 3) or ineffective (fee to fine ratio = 1 : 1) and the treatment. Initial endowments (P1 : P2) for treatment A were $1.10 : $0.60; in treatment B were $1.10 : $0.70; in treatment C were $1.10 : $0.80; in treatment D were $1.10 : $0.90 and in treatment E were $1.10 : $1.10. Thus, if P2 stole $0.20 from P1: in treatment A P1 experienced advantageous inequality ($0.90 : $0.80); in treatment B P1 experienced equal outcomes ($0.90 : $0.90) and in treatments C – E P1 experienced disadvantageous inequality ($0.90 : $1.00, $0.90 : $1.10 & $0.90 : $1.00, respectively). Sample sizes for each condition are indicated in parentheses. Light grey bars, P2 didn't steal; dark grey bars, P2 stole.

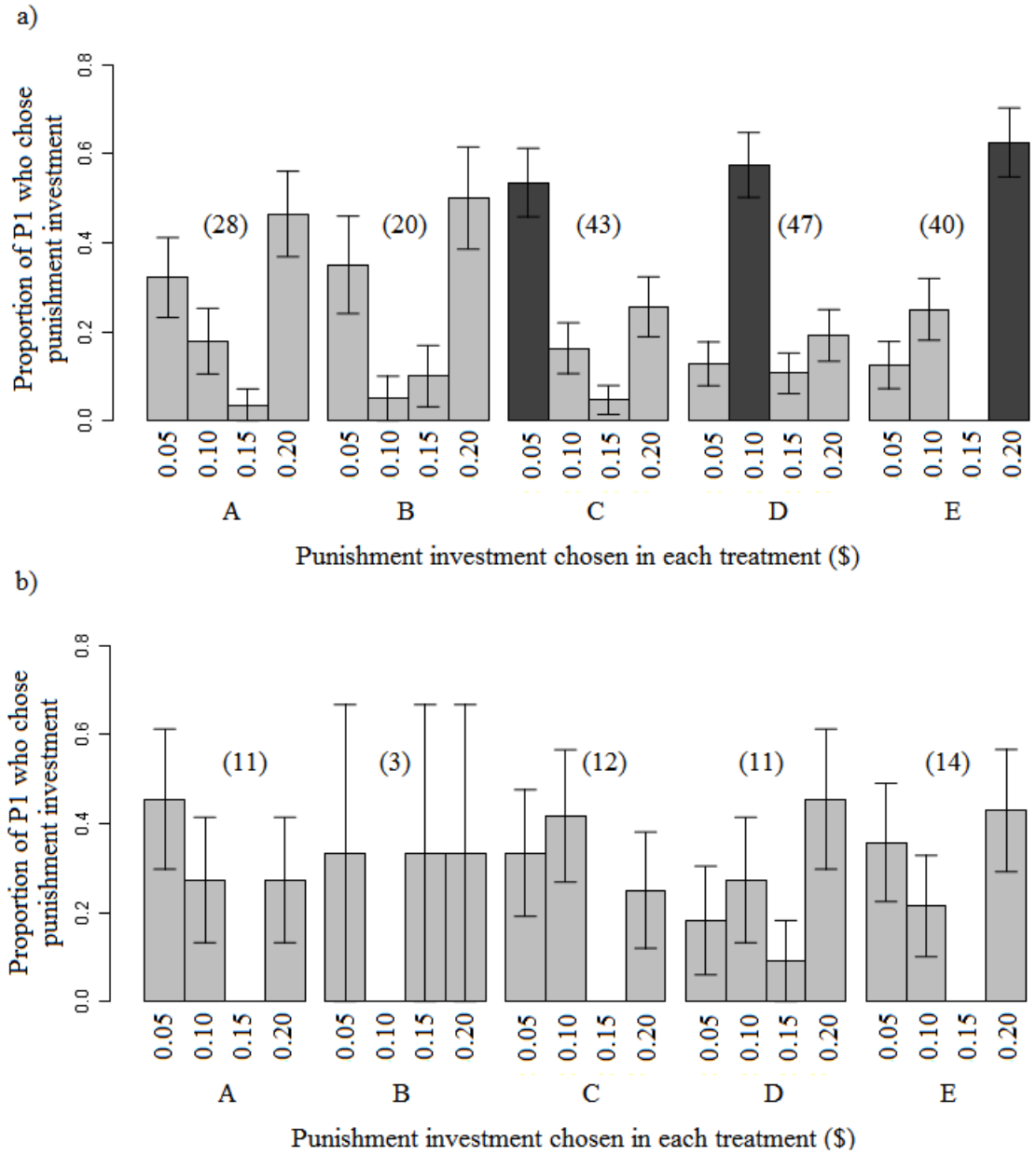

Figure 2. The proportion of punishment investments that were made in treatments A - E (given that P1 punished P2 for stealing) when punishment was **(a)** effective and **(b)** ineffective. If punishment could create equal outcomes, the corresponding punishment investment is shown in dark grey for each treatment; all other punishment investments are shown in light grey. Sample sizes for each treatment are indicated in parentheses.

Treatment	Stage 1 payoff (P1 : P2)	P2 stole (Yes/No)	Stage 2 payoff (P1 : P2)	Outcome (from P1 point of view)	Efficient punishment investment required to create equal outcomes (cost to P1)
A	**$1.10 : $0.60**	**Yes**	**$0.90 : $0.80**	**Advantageous Inequality**	**NA**
		No	**$1.10 : $0.60**	Advantageous Inequality	NA
B	**$1.10 : $0.70**	**Yes**	**$0.90 : $0.90**	**Equal outcomes**	**NA**
		No	$1.10 : $0.70	Advantageous Inequality	NA
C	**$1.10 : $0.80**	**Yes**	**$0.90 : $1.00**	**Disadvantageous Inequality**	**$0.05**
		No	$1.10 : $0.80	Advantageous Inequality	NA
D	**$1.10 : $0.90**	**Yes**	**$0.90 : $1.10**	**Disadvantageous Inequality**	**$0.10**
		No	$1.10 : $0.90	Advantageous Inequality	NA
E	**$1.10 : $1.10**	**Yes**	**$0.90 : $1.30**	**Disadvantageous Inequality**	**$0.20**
		No	$1.10 : $1.10	Equal outcomes	NA

Table 1. The payoffs experienced by P1 and P2 at the beginning of Stage 1 and Stage 2 in treatments A - E. Stage 1 payoffs varied according to the treatment, while Stage 2 payoffs also depend on whether or not P2 stole. These payoffs are described in terms of the outcome (advantageous inequality, equal outcomes or disadvantageous inequality) from P1's point of view. Finally, we show the punishment investment that P1 was required to make to create equal outcomes when punishment was efficient.

Punishment condition	Comparison	P-value	P^{BH}-value	n
Inefficient	P2 didn't steal vs. P2 stole no DI	0.039	0.047	849
	P2 didn't steal vs. P2 stole DI	<0.001	<0.001	975
	P2 stole no DI vs. P2 stole DI	0.073	0.073	618
Efficient	P2 didn't steal vs. P2 stole no DI	<0.001	<0.001	856
	P2 didn't steal vs. P2 stole DI	<0.001	<0.001	987
	P2 stole no DI vs. P2 stole DI	<0.001	<0.001	627

Table 2. The p-values generated by Fisher's exact tests (two-sided) comparing the proportion of P1 that chose a non-zero punishment investment when (i) P2 didn't steal; (ii) P2 stole but the stealing did not result in disadvantageous inequality for P1 ('P2 stole no DI'); and (iii) P2 stole resulting in disadvantageous inequality for P1 ('P2 stole DI'). Comparisons were made for players in both the inefficient and the efficient punishment condition. The fourth column reports Benjamini-Hochberg adjusted p^{BH}-values. The final column shows the sample size (n) for that comparison.