The difference between punishments and rewards in fostering moral concerns in social decision making

Mulder, Laetitia B.

Empfohlene Zitierung / Suggested Citation:
Mulder, L. B. (2008). The difference between punishments and rewards in fostering moral concerns in social decision making. *Journal of Experimental Social Psychology*, *44*(6), 1436-1443. https://doi.org/10.1016/j.jesp.2008.06.004

The Difference between Punishments and Rewards in Fostering Moral Concerns in Social Decision Making

Laetitia B. Mulder

PII: S0022-1031(08)00101-7
DOI: 10.1016/j.jesp.2008.06.004
Reference: YJESP 2103

To appear in: *Journal of Experimental Social Psychology*

Received Date: 31 July 2007
Revised Date: 5 June 2008
Accepted Date: 6 June 2008

Please cite this article as: Mulder, L.B., The Difference between Punishments and Rewards in Fostering Moral Concerns in Social Decision Making, *Journal of Experimental Social Psychology* (2008), doi: 10.1016/j.jesp.2008.06.004

The Difference between Punishments and Rewards in Fostering Moral Concerns in Social Decision Making

Laetitia B. Mulder

Tilburg University

Laetitia B. Mulder, Department of Social Psychology, Center of Justice and Decision Making (JuST), Tilburg University.

The author would like to thank Eric van Dijk for his comments.

Correspondence concerning this article should be addressed to Laetitia Mulder, Department of Economic and Social Psychology, Tilburg University, P.O. Box 90153, 5000 LE Tilburg, The Netherlands. E-mail: l.b.mulder@uvt.nl. Tel. +31 (0) 13 4662483, fax +31 (0) 13 4662067.

Abstract

In social decision making, punishing non-cooperation and rewarding cooperation may not only affect cooperation because of instrumental reasons. They may also evoke moral concerns regarding cooperation as they signal that cooperation is socially approved of and non-cooperation socially disapproved of. I argue that punishments do this to a greater extent than rewards as punishments communicate an obligatory rule and rewards communicate a voluntary rule. Indeed, the first experiment shows that, in a social dilemma, the concept of punishment increased cooperation and the concept of a reward did not. The second experiment showed that participants showed more disapproval towards an offender when there was a punishment for non-compliance than when there was a reward for compliance. These findings suggest that punishing non-cooperation more strongly foster moral concerns regarding cooperation than rewarding cooperation. Possible implications for internalizations are discussed.

Key words: punishments, rewards, moral concerns, moral judgments

The Difference between Punishments and Rewards in Fostering Moral Concerns in Social Decision Making

In modern society authorities regulate various kinds of behavior by using sanctions and rewards. For example, national governments punish speeding on the motorway, and managers try to induce employees to perform well by putting bonuses and promotions in prospect. These external incentives may steer behavior as people will probably want to obtain rewards and avoid punishments. But do these external incentives also affect the *underlying motive* to behave in the desired way? Does, for example, a fine on speeding also increase the feeling that it is morally wrong to speed? Does a bonus on hard working also increase the feeling that it is "a good thing" to make an effort for your company? As an authority who installs an external incentive you may hope that people not merely try to escape a punishment or try to obtain a reward, but also are aware that the undesired behavior is "morally wrong" and the desired behavior is "morally right". After all, when you have convinced them of this, people will, for example, keep to the speed limit even if you are not monitoring them.

In other words, for changing individuals' behavior, individuals should be made aware of the moral norms attached to this behavior. In this paper it is tested how external incentives affect people's moral concerns with regard to the behavior. More specifically, I will test how punishments and rewards differ in this.

The expressive function of punishments

In criminology, one of the most recognized functions of the law in general, and punishment in particular, is general deterrence (Bankston & Cramer, 1974; Bentham, 1970; Williams & Hawkins, 1986). This concept implies that punishments deter

individuals from performing a certain behavior mainly because it makes the behavior less attractive for an individual. However, it is also recognized that the law and punishment systems do more that only making breaking the law less attractive (Cooter, 1998; McAdams, 2000; Williams & Hawkins, 1986). It is argued that punishment systems have an expressive function as they signal what the underlying attitudes in society are and show what is disapproved of in society. In other words, punishment systems provide normative validation of existing norms and they show moral condemnation towards the ones who trespass these norms. Because people are motivated to receive social approval and to avoid social disapproval, a punishment system may in this way indirectly steer people's behavior (McAdams, 2000). In more psychological terms, this means that punishments may encourage *injunctive norms*, which are norms that describe how people should behave (and are normally contrasted against *descriptive norms* which are merely norms that describe how people normally behave). Injunctive norms have been shown to be strong behavioral motivators, even in such a way that they generalize over related behavior in other situations (Cialdini et al., 2006; Reno, Cialdini, & Kallgren, 1993). As punishments seem to have this norm-expresing effect, it is not surprising that in law literature it is posed that they may also influence people's *personal* moral judgments (Cooter, 1998): By expressing what the social norms are, people adhere more strongly to these social norms personally as well and are more likely to judge trespassing these norms as morally wrong.

Also in psychology it is noted that punishments, besides a utilitarian function of deterring people performing a certain behavior, have a second important function, namely retribution. It is found that retribution or "giving just deserts" even forms a more

important motive for people to punish than deterrence (Carlsmith, Darley, & Robinson, 2002; McFatter, 1982). The reason behind this is that trespassing a norm challenges the social norms and moral values within a community and that punishment is a way to reassert these norms and values (Darley & Pittman, 2003; Tyler & Boeckmann, 1997). Punishments thus have a moral connotation in the sense that they are a symbolic means by which people feel that the norms of morality in a society are upheld. Indeed, research has shown that people's desire to punish a norm violator is often driven by moral outrage (Darley & Pittman, 2003).

Thus, the above reasoning suggests that punishment systems are highly associated with moral behavioral norms and that the concept of punishment is strongly connected to our moral sentiments. This may have consequences for the effect that punishment can have on awareness of moral norms. If indeed punishment is associated with moral sentiments, then it may also be the case that the mere installation of a punishment evokes moral concerns (see also Yanagida & Fujii, 2004). Thus, punishing littering or illegal downloading may cause people to realize that these behaviors are immoral and may thus evoke moral disapproval of littering or illegal downloading. Research by Thøgersen (2003) supports this notion. In his study, installing a garbage fee increased people's personal norms against too much garbage production which in turn induced them to decrease their garbage production.

Punishments compared to rewards

Besides punishing undesired behavior, one may try to counter undesired behavior by rewarding the desired behavior instead. When discouraging people to drive in environmentally unfriendly cars or to buy light bulbs that waste energy, one may also

reward people who buy environmentally friendly cars or low-energy light bulbs. The question is whether these rewards equally effect moral concerns compared to punishments.

In this paper it is reasoned that they do not. My reasoning for this is that a punishment communicates a different kind of rule than rewards: Whereas punishments communicate obligatory rules, rewards communicate voluntary rules. The distinction between obligatory rules and voluntary rules (see also O'Gorman, Wilson, & Miller, 2005) parallels the distinction made by Kant (1797/1991) between perfect duties and imperfect duties. Perfect duties are duties you *always* have to perform to be moral (for example, being honest or loyal), and imperfect duties are duties you *sometimes* have to perform to be moral (for example, being charitable or kind). In consequence, people who violate a perfect duty are by definition judged as immoral whereas people who violate an imperfect duty not necessarily. So, an individual can still be moral when (s)he does not give money to a beggar, but not when (s)he tells a lie (Trafimow, Bromgard, Finlay, & Ketelaar, 2005; Trafimow, Reeder, & Bilsing, 2001; Trafimow & Trafimow, 1999). Consequently, violating a perfect duty (i.e. an obligatory rule) is regarded as more "morally wrong" than violating an imperfect duty (i.e. a voluntary rule).

The difference between obligatory behavior and voluntary behavior, may be regarded as a continuum rather than as a dichotomous distinction (see also Trafimow et al., 2005). Consequently, in many moral behaviors, there is some latitude in whether individuals interpret a prescribed behavior as either obligatory or as voluntary. For example, whether being cooperative in social dilemmas is regarded as a moral obligation or as voluntary moral behavior may depend on the specific person (such as social value

orientation, c.f. Beggan, Messick, & Allison, 1988; Liebrand, Jansen, Rijken, & Suhre, 1986; Sattler & Kerr, 1991) or on the context of the specific social dilemma (cf. Pillutla & Chen, 1999).

The presence of either a punishment system or a reward system to promote a certain behavior is likely to be a factor that influences whether people interpret the behavior as obligatory or as voluntary. Because a punishment system punishes people who do not perform the desired behavior, it would communicate that the behavior is obligatory. Because a reward system rewards people who perform the desired behavior, it communicates that the behavior is voluntary. Consequently, by introducing a punishment for undesired behavior, an authority (or whoever introduces the punishment) shows more disapproval of undesired behavior than by introducing a reward for desired behavior and, therefore, communicates to a greater extent that the desired behavior is "morally right" and that the undesired behavior is "morally wrong". Punishments may thus communicate a moral norm and evoke moral concerns to a greater extent than rewards. Consequently, under the presence of a punishment, the behavior will be judged more strongly in moral terms than under presence of a reward. This was tested in two studies. In the first study it was tested in the context of cooperation in a social dilemma. In the second study I focused on the extent to which people show moral disapproval towards someone who does not comply by a rule that is either backed up by a punishment or a reward.

Study 1

Method

Design and participants

In this experiment, 114 first-year undergraduate psychology students from Tilburg University participated (95 female, 19 male, M_{age} = 19 years, SD_{age} = 2.55). The experiment took about 15 minutes and participants received course credits for their participation. They were randomly assigned to one of the three conditions (control, punishment, reward).

Procedure

Upon arrival in the laboratory, participants were seated in separate computer cubicles containing a table, a chair and a computer. Communication with other participants was prohibited. The instructions for the experiment appeared on the computer screen. Participants were told that they were linked to three other research participants in the lab. First they were given the general instruction that they would be presented with a decision situation in which they, similar to the three other group members, were to decide how many "coins" they were going to put in "the pool".

Then, the punishment condition, participants were told the following (reward condition is between parentheses):

> *In decision situations such as these, it often happens that group members choose to punish (reward) those who put only few (put many) coins in the pool. Thus, group members often make other group members who put a few (put many) coins in the pool, pay a fine (obtain a reward). Those who put only few (put many) coins in the pool are thus punished (rewarded).*

It was stressed that, in the situation they were about to face, there would actually be no punishment (or a reward, in the reward condition) because the experimenters wanted to

keep the situation simple. This was specifically done to merely activate the concept of a punishment or a reward without at the same time creating an economical reason to cooperate (i.e. avoiding a punishment or trying to obtain a reward). In this way, cooperation levels would reflect internally rather than external motivated behavior. Nothing was told about a reward or punishment in the control condition.

After this, the situation they faced was explained more thoroughly. Participants were told that they owned 10 coins and so did the other three group members. Each coin was worth 1 Euro (about 1.60 USD). Each group member was given the opportunity to put coins in the pool. Coins that were contributed to the pool would be multiplied by 1.5 and then equally divided among the group members, irrespective of how many coins each group member put in the pool. In this way, the decision situation participants faced was a social dilemma situation as 1) mutual cooperation (putting coins in the pool) yielded more than mutual defection (keeping coins to the self), but 2) keeping the coins for the self yielded more at an individual level than putting coins in the pool (Dawes, 1980). Participants were further told that one of the groups taking part in the experiment would be randomly selected to be actually paid the money they earned in this experiment. To make sure participants understood the situation, they were asked some quiz questions. After each quiz question the correct answer was provided.

Then, moral concerns regarding cooperation were measured with twelve items on a scale ranging from 1 (*totally disagree*) to 7 (*totally agree*). Example items were “It is my moral duty to put coins in the pool”, “Group members who put many coins in the pool are ‘better people’ than group members who put few coins in the pool”, and “I find it is ‘not done’ to put few coins in the pool”. The scale had a high reliability ($\alpha = .88$).

After this, participants were asked how many coins they gave to the pool. Then, to measure their social motives for their decision they were presented four statements (again, on a scale ranging from 1 to 7). These items were: "I wanted to make as much money as possible for myself" (reverse coded), "I wanted to make as much money possible for us together", "I wanted everyone to earn the same amount of money", and "I wanted to earn more than others" (reverse coded). The four items together formed the sufficiently reliable scale "motive of collective-interest" ($\alpha = .60$). Then, the experiment was finished and participants were debriefed. As promised, two weeks later one of the groups was randomly selected and its group members paid according to their decisions in the experiment.

Results

Cooperation

It was hypothesized that the concept of a punishment made participants more cooperative than the concept of a reward. To test this, a one-way ANOVA was performed on the number of coins participants had put into to the pool. It showed that the conditions differed significantly, $F(2,111) = 4.32$, $p < .05$. Participants put more coins in the pool in the punishment condition ($M = 6.76$, $SD = 2.48$) than in the reward condition ($M = 4.97$, $SD = 2.88$) or in the control condition ($M = 5.55$, $SD = 2.63$), LSD Post-hoc analysis, $p < .05$. No difference was found between the reward condition and the control condition..

Moral concerns with regard to cooperation

To test whether moral concerns would be stronger in the punishment condition than in the control and the reward condition, a one-way ANOVA was performed. The results showed a significant effect of the three conditions on moral concerns, $F(2,111) =$

3.12, $p <$. 05. Moral concerns were stronger in the punishment condition (M = 3.86, SD = 1.05) than in both the reward condition (M = 3.34, SD = 0.91) and the control condition (M = 3.36, SD = 1.07), LSD Post-hoc analysis, $p < .05$. No difference was found between the reward condition and the control condition.

Motive of collective-interest

To test whether the punishment had increased the motive of collective interest, a one-way ANOVA was performed. This showed a significant effect of the conditions on the motive of collective interest, $F(2,111) = 5.60$, $p < .005$. The motive of collective interest was stronger in the punishment condition (M = 5.09, SD = 0.90) than in the reward condition (M = 4.48, SD = 1.06) and in the control condition (M = 4.36, SD = 1.10). LSD Post-hoc analysis, $p < .05$. No difference was found between the reward condition and the control condition.

Mediation analyses

Mediation analyses were performed to test whether moral concerns and motive of collective interest respectively mediated the influence of punishment (versus reward) on cooperation. The dummy-coded variable "punishment" was made with value 1 for the punishment condition and value 0 for the control and the reward conditions. The dummy-coded variable "control" was made with the value 1 for the control condition and 0 for the punishment and reward conditions.

First, to contrast punishment against reward, cooperation was regressed on punishment and control. Only punishment predicted cooperation (B = 1.78, $p < .005$). Then, moral concerns were regressed on punishment and control. Again, only punishment was a significant predictor (B = .52, $p < .05$). Finally, when cooperation was regressed on

punishment, control and moral concerns, the effect of punishment on cooperation decreased (B = 1.22, p < .05) and moral concerns predicted cooperation (B = 1.08, p < .001). Thus, moral concerns partially mediated the effect of punishment on cooperation, Sobel's z = 2.00, p < .05.[1]

The same procedure was performed with motive of collective interest as the possible mediator. Punishment significantly predicted motive of collective interest, B = 0.62, p < .05. When regressing cooperation on punishment, control and collective interest, motive of collective interest predicted cooperation (B = 1.41, p < .001) and the effect of punishment on cooperation disappeared (B = 0.92, p = .09). Thus, motive of collective interest fully mediated the effect of punishment on cooperation, Sobel's z = 6.45, p < .001.

So, both moral concerns and motives of collective interest mediated the effect of punishment (versus reward) on cooperation. One could argue that these mediators are correlated and that one of them may be the real mediator that overrules the mediating effect of the other. Indeed, the two measures were positively correlated (r = .38). However, when both motive of collective interest and moral concerns are included in the regression of cooperation on punishment and control, they are both significant predictors (motive of collective interest, B = 1.19, p < .001; moral concerns, B = .68, p < .005).

Discussion

The results of Study 1 confirm the expectations that, when people are confronted with a punishment system, they are more cooperative than when people were confronted with a reward system. It is unlikely that the higher cooperation rates in the punishment condition can be attributed to the fact that people merely wanted to avoid being

sanctioned. After all, instead of a punishment or reward system being actually present, it was merely suggested that either a punishment or reward system in the situation at hand was a common phenomenon. Moreover, the influence of the punishment system on cooperation appeared to be due to the fact that people's moral norm regarding cooperation and people's motive of collective interest was strengthened. This all supports the idea that punishments communicate moral norms of cooperation and encourage moral concerns to a greater extent than rewards, which in turn makes people show more cooperation. The fact that it was the punishment condition and not the reward condition that always differed from the control condition, suggests that punishments made things better rather than that rewards made things worse. This issue will be addressed in the general discussion section of this paper.

Study 2

In the first study, the main focus was on how punishment and rewards affected moral concerns regarding the desired behavior and people's actual behavior. Next, I was interested in the difference between punishment and rewards on people's behavioral reactions to *others* who show undesirable behavior. Research on strong reciprocity shows that people have the tendency to cooperate with others and punish a person who does not cooperate, even if this goes against their direct self-interest (Fehr, Fischbacher, & Gächter, 2002; Gintis, Bowles, Boyd, & Fehr, 2003). Moreover, it also shows that, even when they are an "outsider", people punish others who are non-cooperative in an interaction with someone else, and are even willing to pay for that.

When punishments result in stronger moral concerns than rewards do, it can be expected that punishments, more than rewards, may also result in stronger social

disapproval towards others who show the undesired behavior. Consequently, punishments would lead to more negative behavioral reactions to others who show undesired behavior than to others who show desired behavior. This is what was tested in Study 2. The desired behavior in Study 2 was allocating money in a dictator-type game. It was hypothesized that, in case of a punishment more than in case of a reward, people would be more likely to choose an economically less attractive option over an economically more attractive option in order to put an allocator who shows undesired behavior at a disadvantage and to put an allocator who shows desired behavior at an advantage.

In Study 2 the *actual* presence of a punishment or a reward was varied. In studying a punishment or reward with a specific size, however, it should be taken into account that people's subjective experience of the size of the punishment may differ from their subjective experience of the size of the reward. After all, it is commonly found that there is an asymmetry in the subjective experience of losses and gains in the sense that losses loom larger than gains (Kahneman & Tversky, 1979). Because punishments involve losses and rewards involve gains, participants will probably experience a fine as more severe than a reward of similar size. Because it has also been shown that the severity of a sanction can increase norms of morality (Mulder, Verboon, & De Cremer, 2007), this subjective difference between the size of a punishment and the size of a reward may provide an alternative explanation for why punishments evoke stronger moral norms than rewards. To solve this, individuals' subjective experience of the size of the punishment or reward was measured and controlled for.

Method

In Study 2 participants were told that they took part in an experiment where they were required to evaluate a situation that was happening at that specific time in another experimental session somewhere else in the laboratory.

Design and participants

In this experiment 51 students (18 male, 33 female, M_{age} = 21.1 years, SD_{age} = 2.45) of Leiden University participated. The experiment took 20 minutes. Participants were randomly assigned to either the punishment or reward condition.

Procedure

Participants were seated in experimental cubicles containing a table, a chair and a computer. Communication with other participants was prohibited. The instructions for the experiment appeared on the computer screen.

Participants were led to believe that, in the laboratory, two experiments were running simultaneously: experiment X and Y. They were told that they themselves were taking part in experiment Y and that experiment Y implied judging a situation that took place in experiment X. The situation in experiment X was similar to a dictator game (see also Van Dijk & Vermunt, 2000). In reality, all participants actually participated in experiment Y.

Participants were explained what decision participants faced in experiment X. More specific, participants were told that in experiment X participants formed couples in which one was assigned the role of “allocator” and the other the role of “recipient”. For the allocator and recipient together, 10 chips were available and the allocator could decide how to allocate them. The situation was represented by a picture in such a way that it was clear that neither of the two parties “owned” the chips (yet) and thus that the

way in which the allocator would allocate the chips could not represent the actions "giving away" or "stealing". Allocators knew that, for them, each of the 10 chips was worth 2 EUR (Euros) and that the value of the chips to the recipient would be unknown to the allocators: For the recipient, chips could be worth either less than 2 EUR, more than 2 EUR or equal to 2 EUR. This was done as ambiguity permits multiple interpretations of what is fair (Thompson & Loewenstein, 1992) and in this way participants were given more liberty to judge what would be the most "fair" distribution of chips. By keeping the value of chips for the recipient unknown for the allocator, there was reason for the allocator (and the participants) to believe that another division than a 5/5 division could be seen as more fair than a 5/5 division.

Sanction manipulation

Participants were told that, in experiment X, for each specific allocator/recipient couple, the experimenter would make a request to the allocator *how* to allocate the chips. In the punishment condition, participants were told that experimenter installed a fine with a certain size (that varied between the allocator/recipient couples) that would be imposed on the allocator if (s)he did not follow the experimenter's request. In the reward condition, participants were told that experimenter would give a bonus with a certain size (that varied between the allocator/recipient couples) to the allocator if (s)he *did* follow the experimenter's request. In both conditions it was told that, because the experimenters were only able to check the decisions of some allocators, it was not certain whether allocators who did not follow the request (or: did follow the request in the reward condition) would actually *be* sanctioned (or rewarded in the reward condition): Whether an allocator, was actually fined or rewarded would depend on whether the experimenter

personally monitored the allocation decision of this specific allocator. Further, participants were told that, in experiment X, after the allocator made his or her decision, the computer would *randomly* determine whether the chips would be paid out in *actual* money.

Then, participants were told that they would be informed about the decisions that two allocators in experiment X had made and of which the computer had determined that the chips would not be paid out in actual money. These allocators were referred to as "allocator A" and "allocator B". Participants were told that these two allocators had not been monitored by the experimenter and thus could not be rewarded (or punished) in case that they had (not) followed the experimenter's request. This was to assure participants that both allocators had not actually gained or lost money and that, so far, there was no difference between the allocators in financial outcomes.

Participants were further told that for allocator/recipient couples A and B, the experimenter had requested the allocators to allocate 7 chips to the recipient and 3 chips to themselves. Additionally, the experimenter had communicated to the allocators that the fine they could receive if they allocated less than 7 chips to the recipient (or, in the reward condition, the bonus they could receive if they allocated at least 7 chips to the recipient), was 5 EUR (but recall that this fine or reward was not administered as these specific allocators had not been monitored). To assess whether participants had understood the information about allocator/recipient couples A and B, they were asked two control questions concerning the request of the experimenter and about fine or reward. Four participants answered these questions incorrectly and were left out the analyses.

Then, participants saw on the computer screen that allocator A had followed the request and had allocated 7 chips to the recipient, whereas allocator B had not followed the request and had allocated 5 chips to the recipient.

Dependent variables

Moral judgment about the behavior of allocator B (i.e. the one who had not followed the request) was measured by presenting eleven items on an answering scale ranging from 1 (totally disagree) to 7 (totally agree). The items started with "The fact that allocator B gave less than seven chips to the recipient...." and then continued differently. Three example items are "....is not very social towards recipient B", "...is morally incorrect" and "...I do not judge negatively" (reverse coded). These items formed a highly reliable scale (Cronbach's $\alpha = .95$).

Then, the behavioral measure of putting allocator B at a disadvantage and allocator A at an advantage was measured. This was done by asking participants to choose between two options: The first option was that they could equally split 9 EUR with allocator A (i.e. the allocator who had followed the request) so that both the participant and allocator A would receive 4.5 EUR. If they chose for this option, allocator B would receive nothing. The second option was that they could equally split 10 EUR with allocator B so that both the participant and allocator B would receive 5 EUR. If they chose for this option, allocator A would receive nothing. Participants were asked which of these two options they chose (cf., Kahneman, Knetsch, & Thaler, 1986). From an economic point of view, the second option was the best as it would yield more money (5 EUR) for participants than the first option (4.5 EUR). From a moral point of view, however, participants may have wanted to choose the first option to withhold money

from allocator B (the one who had not followed the request) and to give allocator A (the one who had followed the request) money instead. By using this procedure, it was expected that people would be less likely to choose the more rational second option when they regarded contravening the request of the experimenter as "morally wrong" (and socially disapproved of) and complying with the request of the experimenter as "morally right" (and socially approved of). This is expected in punishment condition.

Then, participants were asked their opinion about the size of the punishment or reward with the statement (text in reward condition is between parentheses) "I think the size of the fine (bonus) that was placed on contravening (complying with) the experimenters' request was...." and was then followed by a seven point answering scale (1 = *too small*, 7 = *too large*).

Finally, participants were fully debriefed about the experiment, thanked and were all paid the equal amount of 3 EUR (Euros). All participants agreed to this procedure.

Results

Judgment of size or punishment or reward

A t-test was performed to test whether participants judged the size of punishment and reward differently. It indeed showed that participants in the punishment condition judged the punishment as larger (M = 4.79, SD = 1.38) than participants in the reward condition judged the reward (M = 3.78, SD = 1.04). So, as expected, the subjective experience of the size of the punishment differed from the subjective size of the reward. The following analyses that are reported are an ANCOVA and a logistic regression with judgment of punishment/reward size as a covariate.

Moral judgment

The hypothesis was that participants would judge the behavior of B (i.e. the allocator who had not followed the request) as "wrong" in moral terms to a greater extent when there was a punishment than when there was a reward. Indeed, participants in the punishment condition had a stronger moral judgment about B's behavior (M = 3.59, SD = 0.28) than participants in the reward condition (M = 2.73, SD = 0.28), $F(1,44) = 4.34$, $p < .05$.

Social disapproval

In addition, it was expected that in the punishment condition participants were more likely to choose the first option (the less economically attractive option of splitting money with allocator A rather than B) than in the reward condition. Indeed, in the punishment condition 54% of the participants chose the first option of not splitting money with allocator B but with allocator A, whereas in the reward condition, only 21% did, B = 1.52, $p < .05$.

Mediation analysis

It was tested whether the effect of punishment versus reward on the behavioral measures of social disapproval was mediated by moral judgment. Punishment versus reward significantly predicted social disapproval (B = 1.52, $p < .05$) and significantly predicted moral judgment (B = -0.86, $p < .05$). When social disapproval was regressed on both punishment versus reward and moral judgment, moral judgment predicted social disapproval (B = -1.23, $p < .001$) and the effect of punishment versus reward disappeared (B = .99, p = .24). This indicated that moral judgment indeed mediated the effect of punishment versus reward on social disapproval (Sobel's z = 1.78, $p < .05$).

Discussion

The results of Study 2 show that the presence of a punishment system, more than the presence of a reward system, causes people to morally disapprove of someone who shows the undesired (either punished or not rewarded) behavior. Consequently, more strongly in presence of a punishment than in presence of a reward, people showed negative behavior towards someone who has shown undesired behavior and positive behavior towards someone who has shown desired behavior, even though this went against their self interest. Again this supports the notion that punishments result in stronger moral concerns than rewards.

General Discussion

The two studies presented in this paper both support the idea that punishments communicate to a greater extent that the undesired behavior is morally wrong (and that the desired behavior morally right) than rewards. The results show that, compared to a reward, a punishment evokes stronger moral concerns and moral judgments, and makes people judge others who show the undesired behavior more negatively. The behavioral implications of this were shown in both studies. In Study 2, people were willing to pay for putting rule violators at a disadvantage compared to rule compliers (so punish rule violators and reward rule compliers). In Study 1, people themselves showed the desired behavior more when they were told that punishing undesired behavior was a common phenomenon than when they were told that rewarding desired behavior was a common phenomenon. This was not only due to stronger moral concerns, but also to a more stronger motive to serve the collective interest rather than the self-interest. So, these findings suggest that punishing non-cooperative behavior (compared to rewarding

cooperative behavior) encourages a norm to be cooperative that is perceived as a moral norm as well.

A question that may be raised is whether punishments, compared to rewards, are better in *increasing* moral concerns or whether they are better in *preventing* moral concerns to *decrease*. Although results of Study 1 suggest that punishments increase moral concerns and rewards fail to do so, I do not wish to jump to conclusions on this subject. The main focus in this paper was on whether punishments affected moral concerns differently than rewards. I will not argue that rewards can never make things worse. In fact, research has shown that rewards may undermine moral concerns (Frey, 1999; Frey & Oberholzer-Gee, 1997; Heyman & Ariely, 2004) or intrinsic motivation (Deci, Koestner, & Ryan, 1999). Neither do I wish to argue that punishments always make things better as research has shown that punishments can sometimes undermine a moral frame as well (Gneezy & Rustichini, 2000; Tenbrunsel & Messick, 1999). But what I *do* argue is that punishments will result in stronger moral concerns than rewards. This may often be the case because punishments increase them more strongly than rewards do. But in some instances punishments may be more successful in preserving moral concerns than rewards. So, the best conclusion to be drawn is that punishments are better than rewards in *fostering* moral concerns.

Disapproval

Previous work by Greitemeyer and colleagues (Greitemeyer & Kazemi, 2008; Greitemeyer & Weiner, 2003, 2006) suggest that rewards lead to more disapproval of morally incorrect behaviors than punishments do. The question may be raised how this relates to the data in the present paper.

Greitemeyer et al.'s research concerned judgments of others that are perceived to have behaved to avoid a punishment or to obtain a reward. Their work shows that judgments about others is influenced by the possibility to attribute their behavior to a punishment or a reward. As avoiding a punishment is seen as a stronger external motivation than striving for a reward, more dispositional attributions about other people's (im)moral behavior are made when they receive a reward for this behavior than when they avoid a punishment. In contrast, the present paper does not concern behavioral attributions to the punishment or reward, but shows that punishments differently affect moral concerns about the behavior *per se*. This finding has implications for both people's own behavior and disapproval of others.

Still, if one *does* consider how possible behavioral attributions may affect disapproval, two predictions can be made from Greitemeyer et al.'s work. First, it is predicted that people who perform moral behavior and get rewarded are judged as more moral than when they perform moral behavior and do not get punished. Second, it is predicted that, when people perform *immoral* behavior *despite* they are punished, they will be judged as *more immoral* than when they perform immoral behavior *despite* they miss out a reward. Although the former prediction is not in line with the idea that punishments result in stronger moral judgments than rewards, the second one is. So, the findings in this paper do not specifically support or contradict Greitemeyer et al.'s work. This is not so surprising considering the different theoretical point made in the present paper, namely that punishments and rewards differently affect moral concerns about the behavior *per se* and, as a result, disapproval of others who show the (un)desired behavior. The data in this paper support this theoretical point.

Punishment, rewards and internalization

The results of the two studies have implications for the differential effect of punishments and rewards on moral motives that *underlie* behavior. Punishment seems to be more successful in fostering a morality based motivation for behavior than rewards. The finding that punishment leads to more cooperation in Study 1 even without a punishment really present, suggests that punishments may be a better way than rewards to internalize a norm. A punishment may not only result in people feeling that certain behavior is morally good or bad, but it may also cause people to stick with the "good" behavior even when they do not actually expect a punishment when showing the bad behavior. This is also in line with research from child psychology by Kochanska and colleagues (Kochanska, Coy, & Murray, 2001; Kochanska, Tjebkes, & Forman, 1998). These authors studied the difference between the "don't" context (in which children are told to refrain from doing something they like) and the "do" context (in which children are told to do something they do not like). Children developed more committed compliance (i.e. compliance when the mother is not there to monitor them) in the "don't" context than in the "do" context. Kochanska and colleagues concluded that telling children what they should *not* do will be internalized to a greater extent than telling children what they *should* do. Their conclusion fits the reasoning that obligatory rules communicate to a greater extent than voluntary rules that not complying by the rule is morally incorrect. As punishments are measures that communicate what one should *not* do and rewards are measures that communicate what one *should* do, punishments may consequently lead to internalization to a greater extent than rewards.

On the other hand, there is also literature that suggest that punishments, compared to rewards, are not a good measure to induce new behavior and internalization. Literature on behavior modification has suggested that, when punishing old behavior, there will still be positive associations with the old behavior, which makes it hard to extinguish. Rewarding new behavior, by contrast, will create new positive associations which can overrule those of the old behavior (Kazdin, 2001; Van der Pligt, Koomen, & Van Harreveld, 2007). In line with this, Skinner (1971, p. 28-29) argued that punishment may merely learn individuals to avoid the punishment, which may imply showing different undesired behaviors rather than showing the desired behavior (see also, Mulder, Van Dijk, De Cremer, & Wilke, 2006). Also, from the area of cognitive dissonance and self-perception theory (Bem, 1972; Festinger, 1957) one may expect people to attribute their behavior to external reasons rather than on internal reasons when an external incentive is present. Similarly, self-determination theory suggest that punishments and rewards harm feelings of autonomy and thus hinder internalization (e.g., Gagné, 2003; e.g., Joussemet, Koestner, Lekes, & Houlfort, 2004). Although there is, as far as I know, no research showing that punishment externalize more than rewards do, one may argue that punishment are more controlling than rewards and restrict freedom to a greater extent (see also Skinner, 1971, p. 28-29). Because of this, they may hinder internalization to a greater extent than rewards.

As stated earlier, I do not want to question whether punishment can undermine intrinsic motivation. But I do suggest that, as far as *moral* behaviors are concerned, this effect may be less strong or even absent for punishment than for rewards. Whereas for the mere purpose of behavioral modification in a non-moral domain rewards may be more

suitable than punishments, in a moral domain, however, something more is going on than merely learning new behavior and unlearning old behavior. In a moral domain people are made conscious of what the *social norm* is. In such a domain, people do not merely act on what behavior yields the most advantages, but especially act on what they think they should do morally and what other people expect them to do. This paper shows that punishments more strongly than rewards communicate the message "behavior X is morally wrong and behavior Y is morally right". As people are motivated to perceive themselves as moral beings (Bandura, 1999; Tsang, 2002) and have the inclination to have a self-serving and egocentric bias when making attributions for their own positive behavior (Miller & Ross, 1975; Ross & Sicoly, 1979) a punishment will not only motivate them to show behavior X, but will also encourage them to attribute their behavior to their moral intentions rather than to the presence of a punishment. So, the extra feature of a punishment that it evokes moral concerns may not only steer people's behavior in the moral direction, but also form a buffer against attributing the externalizing the motivation to choose behavior Y rather than X and thus reduce the danger of undermining intrinsic motivation.

From the above, it follows that a possible boundary condition for a punishment to foster moral concerns and possible internalization of the desired behavior is that this behavior should take place in a moral domain. Additionally, for a punishment to work in the moral domain, it should be clear what behavior is disapproved of and why. Often, sanctions may do this automatically. For example, the most obvious reason why surcharges on the purchase of energy slurping cars are installed is because these cars are an environmental hazard and thus, it is for welfare of the collective better not to use such

cars. In other situations the function of the sanction may be less clear. For example, when there is a fine for returning library books late, it may not be clear that returning library books late is discouraged to counter the harmful consequences to others who want to rent the overdue books and the resulting risk for the library of loosing customers. It may merely be interpreted as additional "administration costs" (see for how fines may induce a business frame of thinking, Fehr & Falk, 2001; Gneezy & Rustichini, 2000). So, punishments may be more likely to trigger moral concerns if the immoral aspects of the punished behavior are already salient to some extent. In situations in which these aspects are not salient, it may be best to provide explicit reasoning of why the punished behavior is wrong. Indeed, research on parents' use of corporal punishment shows that even when this kind of punishment is "backed-up" by reasoning it is highly effective in preventing future misbehavior (Larzelere, 1986; Larzelere & Merenda, 1994).

Another boundary condition of punishment to work better than a reward may be the extent to which it is experienced as *just.* A recognized disadvantage of punishments is that they, because of their coercive nature, evoke more negativity than rewards and may be regarded as unjust (Depoorter & Vanneste, 2005). For example, corporal punishment may be a particular kind of punishment that is perceived as unjust and harms the attachment bond between parents and children. Indeed, high levels of corporal punishment are related to aggression and lower levels of moral internalization (Gershoff, 2002). However, corporal punishment is not shown to be ineffective when it is used mildly and non-predominantly (Larzelere & Kuhn, 2005). Also, research in social decision making by Van Prooijen, Gallucci and Toeset (in press) shows that only when (financial) punishment is unfair, it can decrease compliance. So, it may be especially in

cases in which punishment is regarded as just, for example when the authority who introduces them is perceived as trustworthy and legitimate, and when the punishment is not too harsh, that punishments evoke a moral norm more strongly that rewards do.

Practical implications

In many real life situations, authorities may only be interested in changing people's behavior. For this particular purpose, rewards may be better than punishments. However, often it is desired that people not only change their behavior but also that they become aware that the new behavior is "morally correct" and the old behavior "morally incorrect". Punishments seem to be better able to reach this goal than rewards. Of course, an ideal way to change behavior in a moral domain would be to *combine* rewarding the desired behavior and punishing the undesired behavior. However, when this is too complicated or expensive, one may well be advised to install a punishment rather than a reward because it more strongly communicates as moral message. So, if authorities, for example, are concerned with fostering a moral norm among citizens that one should stop buying energy slurping cars and buy energy-conserving cars instead, they should *punish* buying energy-slurping cars rather than rewarding buying environmentally friendly cars. This will not only foster moral concerns, but will also make it more likely that citizens will call each other to account when buying energy-conserving cars. Possibly, a more general norm may then be encouraged in favor of considering the environment in everyday choices, affecting other environment-related behaviors.

Concluding remarks

The present paper shows that punishment foster moral concerns to a greater extent than rewards, either because of increasing moral concerns (as Study 1 suggests) or

preventing a decrease in moral concerns. The extent to which punishment (compared to rewards) encourage internalization is still a topic of discussion. Two proposed boundary conditions are that the behavior should be in a moral domain and that the punishment should be regarded as just. The exact effect of punishments compared to rewards on internalization is an interesting topic for future research and should best be studied longitudinally. From the present findings can be concluded, however, that as punishments foster people's moral concerns more than rewards, they are more likely to affect behavior in a positive way even if people expect not to get punished (or rewarded) for real (for example, when behavior is not perfectly monitored, or when the punishment or reward system is removed). As punishment, more than rewards, evokes social disapproval of showing the undesired behavior, it may not only be a better method to foster moral concerns with regard to individual behavior, but also a better method to uphold a norm within a social community. Of course, this is the best way to make people behave accordingly.

References

Bandura, A. (1999). Moral disengagement in the perpetration of inhumanities. *Personality and Social Psychology Review, 3*, 193-209.

Bankston, W. B., & Cramer, J. A. (1974). Toward a macro-sociological interpretation of general deterrence. *Criminology, 12*, 251-280.

Beggan, J. K., Messick, D. M., & Allison, S. T. (1988). Social values and egocentric bias: Two tests of the might over morality hypothesis. *Journal of Personality and Social Psychology, 55*, 606-611.

Bem, D. J. (1972). Self-perception theory. In L. Berkowitz (Ed.), *Advances in Experimental Social Psychology* (Vol. 6). New York: Academic Press.

Bentham, J. (1970). *An introduction to the principles of morals and legislation*. London: Athlone.

Carlsmith, K. M., Darley, J. M., & Robinson, P. H. (2002). Why do we punish? Deterrence and just deserts as motives for punishment. *Journal of Personality and Social Psychology, 83*, 284-299.

Cialdini, R. B., Demaine, L., Sagarin, B. J., Barret, D. W., Rhoads, K., & Winter, P. L. (2006). Managing social norms for persuasive impact. *Social Influence, 1*, 3-15.

Cooter, R. (1998). Expressive law and economics. *Journal of Legal Studies, 27*, 585-608.

Darley, J. M., & Pittman, T. S. (2003). The psychology of compensatory and retributive justice. *Personality and Social Psychology Review, 7*, 324-336.

Dawes, R. M. (1980). Social dilemmas. *Annual Review of Psychology, 31*, 169-193.

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

Depoorter, B., & Vanneste, S. (2005). Norms and enforcement: the case against copyright law. *Oregon Law Review, 84*, 1127-1179.

Fehr, E., & Falk, A. (2001). Psychological foundations of incentives. *European Economic Review, 46*, 687-724.

Fehr, E., Fischbacher, U., & Gächter, S. (2002). Strong reciprocity, human cooperation, and the enforcement of social norms. *Human Nature, 13*, 1-25.

Festinger, L. (1957). *A theory of cognitive dissonance*. Oxford, England: Row, Peterson.

Frey, B. S. (1999). Morality and rationality in environmental policy. *Journal of Consumer Policy, 22*, 395-417.

Frey, B. S., & Oberholzer-Gee, F. (1997). The cost of price incentives: An empirical analysis of motivation crowding-out. *The American Economic Review, 87*, 746-755.

Gagné, M. (2003). The role of autonomy support and autonomy orientation in prosocial behavior engagement. *Motivation and Emotion, 27*, 199-223.

Gershoff, E. T. (2002). Corporal punishment by parents and associated child behavior and experiences: A meta-analytic and theoretical review. *Psychological Bulletin, 128*, 539-579.

Gintis, H., Bowles, S., Boyd, R., & Fehr, E. (2003). Explaining altruistic behavior in humans. *Evolution and Human Behavior, 24*, 153-172.

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

Greitemeyer, T., & Kazemi, E. (2008). Asymmetrical consequences of behavioral change through reward and punishment. *European Journal of Social Psychology, 38*, 246-259.

Greitemeyer, T., & Weiner, B. (2003). Asymmetrical attributions for approach versus avoidance behavior. *Personality and Social Psychology Bulletin, 29*, 1371-1382.

Greitemeyer, T., & Weiner, B. (2006). Effects of reward and coercion for transgression compliance on judgments of responsibility and recommended punishment. *Journal of Applied Social Psychology, 36*, 3015-3030.

Heyman, J., & Ariely, D. (2004). Effort for payment. A tale of two markets. *Psychological Science, 15*, 787-793.

Joussemet, M., Koestner, R., Lekes, N., & Houlfort, N. (2004). Introducing uninteresting tasks to children:A comparison of the effects of rewards and autonomy support. *Journal of Personality, 72*, 139-166.

Kahneman, D., Knetsch, J. L., & Thaler, R. H. (1986). Fairness and the assumptions of economics. *Journal of Business, 59*, S285-S300.

Kahneman, D., & Tversky, A. (1979). Prospect theory: an analysis of decision under risk. *Econometrica, 47*, 263-291.

Kant, I. (1797/1991). *The metaphysics of morals*. Cambridge, UK: Cambridge University Press.

Kazdin, A. E. (2001). *Behavior modification in applied settings*. Belmont, CA: Wadsworth.

Kochanska, G., Coy, K. C., & Murray, K., T. (2001). The development of self-regulation in the first four years of life. *Child Development, 72*, 1091-1111.

Kochanska, G., Tjebkes, T. L., & Forman, D. R. (1998). Children's emerging regulation of conduct: Restraint, compliance, and internalization from infancy to the second year. *Child Development, 69*, 1378-1389.

Larzelere, R. E. (1986). Moderate spanking: Model or deterrent of children's aggression in the family? *Journal of Family Violence, 1*, 27-36.

Larzelere, R. E., & Kuhn, B. R. (2005). Comparing child outcomes of physical punishment and alternative disciplinary tactics: A meta-analysis. *Clinical Child and Family Psychology Review, 8*, 1-37.

Larzelere, R. E., & Merenda, J. A. (1994). The effectiveness of parental discipline for toddler misbehavior at different levels of child distress. *Family Relations, 43*, 480-488.

Liebrand, W. B., Jansen, R. W. T. L., Rijken, V. M., & Suhre, C. M. J. (1986). Might over morality: Social values and the perception of other players in experimental games. *Journal of Experimental Social Psychology, 22*, 203-215.

McAdams, R. H. (2000). An attitudinal theory of expressive law. *Oregon Law Review, 79*, 339.

McFatter, R. (1982). Purpose of punishment: Effects of utilities of criminal sanctions on perceived appropriateness. *Journal of Applied Psychology, 67*, 255-267.

McKinnon, D. P., Fairchild, A. J., & Fritz, M. (2007). Mediation analysis. *Annual Review of Psychology, 58*, 593-614.

McKinnon, D. P., Lockwood, C. M., Hoffman, J. M., West, S. G., & Sheets, V. (2002). A comparison of Methods to test mediation and other intervening variable effects. *Psychological Methods, 7*, 83-104.

Miller, D. T., & Ross, M. (1975). Self-serving biases in the attribution of causality: Fact or fiction? *Psychological Bulletin, 82*, 213-225.

Mulder, L. B., Van Dijk, E., De Cremer, D., & Wilke, H. A. M. (2006). When sanctions fail to increase cooperation in social dilemmas: Considering the presence of an alternative defection option. *Personality and Social Psychology Bulletin, 32*, 1312-1324.

Mulder, L. B., Verboon, P., & De Cremer, D. (2007). Sanctions and moral judgments: The moderating effect of sanction severity and trust in authorities. *Manuscript submitted for publication*.

O'Gorman, R., Wilson, D. S., & Miller, R. R. (2005). Altruistic punishment and helping differ in sensitivity to relatedness, friendship, and future interactions. *Evolution and Human Behavior, 26*, 375-387.

Pillutla, M. M., & Chen, X. P. (1999). Social norms and cooperation in social dilemmas: The effects of context and feedback. *Organizational Behavior and Human Decision Processes, 78*, 81-103.

Reno, R. R., Cialdini, R. B., & Kallgren, C. A. (1993). The transsituational influence of social norms. *Journal of Personality and Social Psychology, 64*, 104-112.

Ross, M., & Sicoly, F. (1979). Egocentric biases in availability and attribution. *Journal of Personality and Social Psychology, 37*, 322-336.

Sattler, D. N., & Kerr, N. L. (1991). Might versus morality explored: Motivational and cognitive bases for social motives. *Journal of Personality and Social Psychology, 60*, 756-765.

Skinner, B. F. (1971). *Beyond freedom and dignity*. New York: Alfred A. Knopf, inc.

Tenbrunsel, A. E., & Messick, D. M. (1999). Sanctioning systems, decision frames, and cooperation. *Administrative Science Quarterly, 44*, 684-707.

Thøgersen, J. (2003). Monetary incentives and recycling: Behavioural and psychological reactions to a performance-dependent garbage fee. *Journal of Consumer Policy, 26*, 197-228.

Thompson, L., & Loewenstein, G. (1992). Egocentric interpretations of fairness and interpersonal conflict. *Organizational Behavior and Human Decision Processes, 51*, 176-197.

Trafimow, D., Bromgard, I. K., Finlay, K. A., & Ketelaar, T. (2005). The role of affect in determining the attributional weight of immoral behaviors. *Personality and Social Psychology Bulletin, 31*, 935-948.

Trafimow, D., Reeder, G. D., & Bilsing, L. M. (2001). Everybody is doing it: the effects of base rate information on correspondent inferences from violations of perfect and imperfect duties. *The Social Science Journal, 38*, 421–433.

Trafimow, D., & Trafimow, S. (1999). Mapping perfect and imperfect duties onto hierarchically and partially restrictive trait dimensions. *Personality and Social Psychology Bulletin, 25*, 687-697.

Tsang, J. A. (2002). Moral rationalization and the integration of situational factors and psychological processes in immoral behavior. *Review of General Psychology, 6*, 25-50.

Tyler, T. R., & Boeckmann, R. J. (1997). Three strikes and you are out, but why? The psychology of public support for punishing rule breakers. *Law and Society Review, 31*, 237-265.

Van der Pligt, J., Koomen, W., & Van Harreveld, F. (2007). *Bestraffen, belonen en beïnvloeden. Een gedragswetenschappelijk perspectief op handhaving [Punishing, rewarding and influencing. A behavior scientific perspective on enforcement].* The Hague, The Netherlands: Boom Juridische uitgevers.

Van Dijk, E., & Vermunt, R. (2000). Strategy and fairness in social decision making: Sometimes it pays to be powerless. *Journal of Experimental Social Psychology, 36*, 1-25.

Van Prooijen, J. W., Gallucci, M., & Toeset, G. (in press). Procedural justice in punishment systems: Inconsistent punishment procedures have detrimental effects on cooperation. *British Journal of Social Psychology*.

Williams, K. R., & Hawkins, R. (1986). Perceptual research on general deterrence: A critical review. *Law and Society Review, 20*, 545-572.

Yanagida, M., & Fujii, S. (2004). The psychological effects of institution of law on moral obligation. In *Proceedings of Master Thesis of Department of Civil engineering*. Tokyo: Tokyo Institute of Technology.

Footnotes

[1] Because the p-values that Sobel tests report are regarded as too conservative (McKinnon, Fairchild, & Fritz, 2007; McKinnon, Lockwood, Hoffman, West, & Sheets, 2002) I used, for all Sobel tests reported in this paper, the corrected p-values as provided by McKinnon et al. (McKinnon et al., 2002).