The importance of unfair intentions and outcome inequality for punishment by third parties and victims

Stefanie Hechler* & Thomas Kessler

Friedrich Schiller University of Jena, Germany

Author note

*Correspondence concerning this article should be addressed to Stefanie Hechler

Author information: Stefanie Hechler, Friedrich-Schiller-University Jena, Institute of Psychology, Department of Social Psychology, Humboldtstraße 26, 07743 Jena, Germany. E-mail: stefanie.hechler@uni-jena.de. Phone: +49- 3641- 945256.**;** Thomas Kessler, Friedrich-Schiller-University Jena, Institute of Psychology, Department of Social Psychology, Humboldtstraße 26, 07743 Jena, Germany. E-mail: Thomas.kessler@uni-jena.de. Phone: +49-3641- 945254.

Author contribution: All authors contributed equally to the conceptualization of the study. Stefanie Hechler drafted the manuscript. Thomas Kessler provided comments and revisions. We thank Julia Elad-Strenger and two anonymous reviewers for comments on an earlier draft.

Statement regarding ethics: The Ethics Commission of the Faculty of Psychology of the Friedrich Schiller University of Jena will give approval to conduct the present experiment.

Abstract

Retributive theories often focus on observers' motives for punishment, which are affected more by the offender's malicious intentions, than the actual outcome of the offense. However, victims experience an offense from a different perspective. The value/status approach argues that an offense has two facets that produce different threats: the intentional violation of values and status imbalance between offender and victims. Whereas third parties and victims punish unfair intentions, victims may also punish the resulting negative outcome inequality between them and the offender. In the proposed study, we orthogonally cross the factors offender's intention with the actual outcome. The participants are either in the role of a third party or of the victim and decide on the punishment of the offender. This approach captures qualitative differences of third-party punishment and punishment by victims. We will discuss how the findings relate to retributivism and other psychological theories of punishment.

The importance of unfair intentions and outcome inequality for punishment by third parties and victims

Introduction

Problem

Individuals' intuitions to punish are generally retributive. They punish offenders proportional to the severity of the offense without much further regard of whether the punishment may affect deterrence or rehabilitation of an offender (Carlsmith & Darley, 2008; Gromet & Darley, 2009a; Vidmar, 2001). Considering the severity of the deed as the main determinant of punishment may miss the different perspective of victims and third parties, because it comprises two conceptually different facets of an offense: the offender's malicious intentions and the actual negative outcome.

The value/status approach argues that these two facets produce conceptually different threats (Wenzel, Okimoto, Feather, & Platow, 2008): First, an offense raises doubts about the validity and importance of the violated norms and values, because the offender intentionally violated them. Second, an offense produces a status imbalance between offender and victim, because it benefits the offender at the expense of the victim. Considering the value/status approach, we suggest that although third parties and victims similarly punish for offender's *malicious intentions*, such as unfair decisions, they differ in their punishment tendencies of the mere *inequality* produced by the offense. The proposed experiment attempts to disentangle the influence of these two theoretically distinct facets of an offense and demonstrate different punishment motives in third parties and victims.

Existing research focuses on the punishment by either victims or third parties (i.e., observers) but a systematic comparison between both is rare (e.g., see Gummerum & Chu, 2014). If victims and third parties follow different intuitions about what to punish then it would be difficult to create punishment that is perceived as just by all. This could threaten

alternative approaches to conflict resolution, which include victims as well as the community (i.e., third parties) in the justice processes. Therefore, it is important to investigate whether the punishment intuitions of victims differ from the intuitions of third parties.

Review of relevant scholarship

Retribution as psychological motive for punishment

In philosophical research retributivism received several (normative) conceptualisations (e.g., Cottingham, 1979; Walker, 1999). However, most scholars agree that the crime suffices to justify punishment (see for example, Moore, 1993). According to Kant (1785 /2007), perpetrators have to be punished because of their 'internal wickedness', which is expressed by the crime committed. Other potential effects of punishment are not central for its justification such as reform of the offender or deterrence of the offender or others of future crimes. Hence, the maliciousness of the offender' intention, and thereby the offenders' responsibility and guilt determine 'just' punishment.

Psychological studies indicate that lay intuitions about punishment correspond to retributivism (Carlsmith, Darley, & Robinson, 2002; Gromet & Darley, 2009b). According to these studies, punishment is thought to re-establishes the moral imbalance produced by the offense (Bies & Tripp, 1996; Carlsmith & Darley, 2008; Darley & Pittman, 2003). Therefore, a just punishment corresponds to the severity of the offense. Indeed, punishers consider the severity of the offense to determine the amount of punishment (for comprehensible overviews see: Carlsmith & Darley, 2008; Gromet & Darley, 2009b). Generally, individuals punish more severe offenses (e.g., murder) stronger than less severe offenses (i.e., theft; Gromet & Darley, 2006). The perceived severity seems to imply the offender's intention. More severe offenses require more malicious intentions, such as the intention to kill or to steal. Moreover, intentionally produced harm is judged more severely and punished more severely than unintentional harm (Ames & Fiske, 2013; Darley & Huff, 1990). Thus, the intention and the negative outcome of an offense are closely associated, even though they differ conceptually.

Their separate effects are not assessed when focusing on the severity of the deed as determinant for punishment. Moreover, participants in these studies are usually in a third-party perspective, which does not reveal potential differences in the perception of victims and third parties.

Intention, value violation and punishment

To understand the influence of malicious intentions and the produced negative outcome on punishment, one has to differentiate between the intention to do something and the actual outcome (Hechler & Kessler, 2018; Malle & Knobe, 1997). This notion is reflected in the value/status approach to punishment (Wenzel et al., 2008). According to the value/status approach, an offense produces at least two conceptually different threats: an intentional value violation and status imbalance between offender and victim. Whereas a value violation implies malicious intentions of the offender, the negative outcome reflects a status imbalance between offender and victim.

An intentional deviation from "how one ought to behave" given the moral convictions, perceived norms, expectations, and/or codes of conducts (here referred to as values) within a particular context threatens their validity (Fiske & Tetlock, 1997; Hechler, Neyer, & Kessler, 2016; Mendoza, Lane, & Amodio, 2014; Okimoto & Wenzel, 2010). Respectively, an offense is seen as a value violation when the offender acts intentionally: someone intends to produce a negative outcome for a victim in an act that violates values that are known to the actor (Cushman, 2008; Gray, Young, & Waytz, 2012; Mikhail, 2007). Actions that are beyond the control of an individual are seen as accidents. Whereas accidents can also produce negative outcomes (e.g., a car accident with fatalities), they do not violate values.

With punishment, people express their disapproval, and moral outrage of the action, and thereby distance themselves from attempted or actual offense and the offender (Eidelman & Biernat, 2003; Feinberg, 1965). The punishment of ingroup offenders, for instance, promotes a positive group identity (Hutchison, Abrams, Gutierrez, & Viki, 2008) and

punishment increases the punisher's moral standing (Hofmann, Brandt, Wisneski, Rockenbach, & Skitka, 2018). This is particularly true, when the punishers expect that the offenders know the values, for example when they share a common identity (Pinto, Marques, Levine, & Abrams, 2010; Shinada, Yamagishi, & Ohmura, 2004). Mitigating factors are to be new in a group (i.e., not knowing the values) or ambiguous intentions, which elicit less desire to punish (Otten & Gordijn, 2014; van Prooijen, 2006).

For third parties (at least those unrelated to the victims), the perceived value violation is the primary cause for punitive tendencies (Okimoto & Wenzel, 2010; Tyler & Boeckmann, 1997). In line with this reasoning, unambiguous malicious intentions (and with that a violation of their values) determine moral condemnation, outrage and punishment in third parties (Gray, Young, & Waytz, 2012; Mikhail, 2007; van Prooijen, 2006). This holds for successful actions that produce negative outcomes (e.g., murder) as well as for unsuccessful acts that do not produce such outcomes (e.g., attempted murder; Cushman, 2008; Hechler & Kessler, 2018). In contrast, accidental or non-intentional negative outcomes do not elicit moral condemnation, outrage or punishment in third parties. Respectively, third parties invest more of their own resources to punish offenders who intentionally chose the less equal distribution in contrast to those who chose the more equal distribution (Gummerum & Chu, 2014).

Victims also often share a group identity with offenders or believe that offenders should appraise the same values as them. Accordingly, they punish those they expect to adhere to fairness norms and intentionally violate them. For example, they reject unfair ingroup offers more often than unfair outgroup offers (Mendoza et al., 2014). Moreover, victims punish identical offers more if the offender intentionally chose the more unequal outcome over the less unequal outcome (Falk, Fehr, & Fischbacher, 2003; Gummerum & Chu, 2014). Similarly, the same unequal distribution elicits more punishment when offered by

a person (i.e., intentional offender) than by a random device (Blount, 1995; Falk, Fehr, & Fischbacher, 2008).

Outcome equality, status imbalance and punishment

In addition to the threat to values through intentional violations, offenses produce a status imbalance between offenders and the victims (Wenzel et al., 2008). An offense reveals that the victim has less control over the situation compared to the offender (SimanTov-Nachlieli, Shnabel, & Nadler, 2013), and as a consequence may be harmed or left with fewer resources. However, such status imbalance can also emerge without explicit intentions from the offender, as research on relative deprivation indicates (e.g., Kessler & Mummendey, 2002; Smith, Pettigrew, Pippin, & Bialosiewicz, 2012).

Status imbalance can elicit willingness to harm or to punish the advantaged, in particular in competitive relations and when norms of fairness and cooperation are absent (Fehr & Schmidt, 1999; Raihani & Bshary, 2019). When cooperative norms are absent individuals are even willing to harm cooperating group members (socalled antisocial punishment; Herrmann, Thöni, & Gächter, 2008). Status imbalances increase the salience of interpersonal differences for victims relative to a common social identity, which enhances antagonism and harming of the advantaged (Skitka, 2003; Turner, Oakes, Haslam, & McGarty, 1994). In intergroup relations, groups perceive an attack by an outgroup rather as a status threat than a value violation (Okimoto & Wenzel, 2010). Thus, victims are likely to perceive their relationship with offenders as antagonistic and competitive.

In competitive relations, being disadvantaged can elicit willingness to harm or to punish the advantaged even if a value violation by an offender is absent (Fehr & Schmidt, 1999; for a review, see: Raihani & Bshary, 2019). Previous studies have shown that victims reduce the offender's payoff more because of the inequality between them and the offender than because of the absolute losses they suffered (Raihani & McAuliffe, 2012). Moreover, victims who obtain negatively interdependent relationships with offenders punish inequality

more than independent victims (Marczyk, 2017). Nevertheless, victims do not necessarily aim at decreasing outcome inequality between them and the offender. They punish offenders or reject unequal offers even when it has no effect on the inequality (Bone & Raihani, 2015; Falk, Fehr, & Fischbacher, 2005). Thus, in status imbalanced contexts, victims are motived to punish inequality.

Even though third parties primarily punish for offender's intentions and not accidents they sometimes also punish status imbalances. This action on behalf of the victims is found when third parties care for them, empathize, or share a common identity with the victims (Batson, Chao, & Givens, 2009; Bernhard, Fischbacher, & Fehr, 2006; Gromet, Okimoto, Wenzel, & Darley, 2012; Lieberman & Linke, 2007; Pfattheicher, Sassenrath, & Keller, 2019). Such empathic action may include punishment of the advantaged or compensation of the victims. Although it is not clear whether punishment or compensation dominates (Chavez & Bicchieri, 2013; Van Prooijen, 2010), studies have shown that third parties with unspecific anger tend to punish whereas those primed with empathic anger prefer compensation (Gummerum, Van Dillen, Van Dijk, & López-Pérez, 2016). Thus, third parties are less concerned about status imbalances between offenders and victims. Only when they relate to the victims they act on their behalf.

Hypothesis, aims and objectives

Most offenses imply intentional value violations and produce status imbalance at the same time. These facets are differentially important to third parties and victims. Therefore, these two facets must be disentangled carefully to test their specific effects on punishment in a full experimental design. According to the value/status approach, third parties primarily punish value violations that imply offender's intention, whereas victims also punish for status imbalance such as outcome inequality. The proposed study replicates findings that third parties punish malicious intentions, but not unintentional negative outcomes (Cushman, 2008; Gummerum & Chu, 2014; Hechler & Kessler, 2018). Moreover, it replicates prior studies that

tested how victims punish offenders who intentionally distributed unfairly or fairly, and who had full or limited influence on the outcome equality (Falk et al., 2008; Gummerum & Chu, 2014). To our knowledge, there has not yet been a study that directly compared punishment by third parties and victims in respect to offender's intention and outcome inequality.

Specifically, we propose that (H1) offender's intention (unfair decision vs. fair decision) elicits punishment in both third parties and victims, whereas (H2) actual outcome (inequality vs. equality) elicits more punishment in victims than in third parties. Hence, we expect a main effect of offender's intentions on punishment, and a two-way interaction of punisher's perspective (victim/third party) and outcome inequality on punishment, but no two-way interaction of punisher's perspective and intentions on punishment. These hypotheses are presented in Figure 1.

Figure 1. Expected results according to H1 and H2

Design Plan

The following experiment orthogonally manipulates the factors offender intention, outcome, and participants' perspective and measures as a dependent variable punishment of the offender. Accordingly, the experiment will have a three-factorial (2×2×2) design with the within factors *intention* (unfair/fair) and *outcome* (unequal/equal), and the between factor *perspective* (third party/ victim).

Materials and Methods

Data collection will be conducted online by the ZDIP's PsychLab, in Summer 2020.[1] Participants will receive compensation of at least 1.25 € plus their incentives from the critical rounds (maximum 1.5 €) from the game. Participation will take approximately 15 minutes. The study will be programmed on the online platform Soscisurvey (Leiner, 2018), which participants will enter using a link provided in the invitation. Each participant is randomly assigned to one of the two *perspective* conditions (third-party or victim) using a full randomization procedure. Instructions in each condition merely differ according to perspective.

Ethical approval will be obtained from the local ethic-committee at the University of Jena and informed consent of participants. The study will be preregistered as part of the current Special Issue. All study materials, measures, (fully anonymized) data sets and analysis scripts in R will be openly shared on PsychArchives and/or Open Science Framework.

Sample size, power and precision

We will conduct a three-way factorial ANOVA to examine the effects of the manipulated factors on punishment. The sample size was determined using G*Power to calculate the a-priori effect size needed to detect main effects (H1) and two-way interactions (H2) with high (95%) power (Faul, Erdfelder, Lang, & Buchner, 2007). Gummerum and colleagues (2014) investigated victims and third-party punishment about offender's decisions on more and less equal outcomes varying from advantageous (2/8) to disadvantageous distributions (10/0) to the victim. In adults, they found a small effect of decision on victims ($\eta^2 = 0.08$), as well as on third parties ($\eta^2 = 0.06$) using a 1x4 design for each group. We use this as the expected effect size to approach H1. A power analysis with $\eta^2 = .06$ indicated that we would need a total sample size of 53. No prior studies – to our knowledge – directly compare how much victims and third parties punish inequality. Marczyk (2017) shows that

[1] Since we are currently recommended to avoid face-to-face interaction due to the spread of the Corona-Virus, we decided for online data collection. We will use a laboratory setting if feasible and supported by the ZPID.

participants more often punished a decision for inequality more, if it was at the expense of the participant (interdependence, similar to victim perspective; 75%) than when the outcomes of the players were independent (similar to third party perspective; 18%). A decision for an equal outcome was still punished more in the interdependent (47%) than the independent condition (14%). Different from our study, the dependent variable was a dichotomous choice and the interdependent conditions involved unfair intention, whereas the independent did not. Nevertheless, based on these considerations, we assume a small effect of the perspective x outcome interaction on punishment (H2; f= .10). This also provides the opportunity to detect small effects of other two-way interactions. We calculated the a priori sample size using a mixed ANOVA with two groups and two measurements. The data of 328 participants allows to detect a small effect with a power of $1 - ß$ = .95 and at a significance level of 5%. We aim at collecting data of 353 participants to compensate for exclusion of participants (see participant characteristics).

Participant characteristics

The sample will consist of participants from the ZDIP's PsychLab participant pool, who speak German as their mother tongue. Half of them will be female, and they all will be adults (older than 18 years). The phenomenon of costly punishment has been observed across various populations (e.g., Henrich et al., 2006). Thus, a more heterogeneous sample should not change the effects.

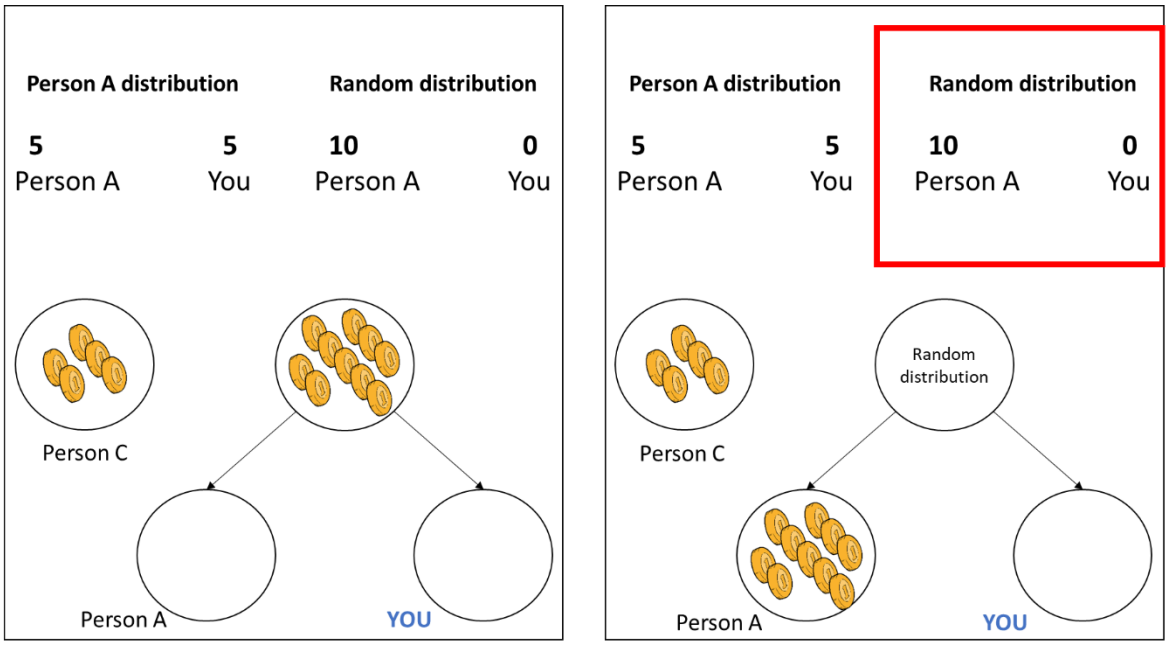
We will exclude participants from data analysis based on processing time and an attention check (i.e., a lure question). We assume that our sample will provide similar properties as an MTurk sample, such as participating anonymously online for a small participation fee. MTurkers have been found to fail lure questions with a probability of 7% when the question is raised in the beginning (see Hauser & Schwarz, 2016, Study 1). Thus, we decide to recruit 353 participants to account for potential exclusions.

Procedure

Participants will be invited via e-mail to participate in an online study on resource allocation behaviour. Upon clicking on a link, participants read that they will interact with others in allocation tasks. After signing an informed consent form, they read the task instructions. In the following interactions, they can collect monetary units (MUs). Each MU has the value of 5 Eurocents. Participants receive their total payoff at the end of the study.

They further read that they are grouped together with allegedly two other participants in each interaction. Participants will be informed of three possible roles in the game:, the decision-maker - Person A, the recipient (i.e. victim) - Person B, and the third party - Person C. In actuality, participants will be randomly assigned to one of two roles, person B or C, and maintain it throughout the experiment. In each round, 10 monetary units (MU) are distributed between person A and B. Person C receives 5 MUs. Person A decides how many of the 10 MUs she would keep and how many she would give to Person B. Simultaneously, a computerized random device distributes the 10 MUs between Person A and Person B. A dice roll puts one of the distributions into place: person A's decision or the random distribution. Person A and Person B receive the MUs based on the dice roll (see Figure 2). Person C observes the interaction. Thus, the decision-maker suggests a distribution, on which the victim has no influence. The outcomes of the decision-maker and the victim are negatively interdependent, whereas the third party's outcome is independent of the others'.

Figure 2. Visual display of one interaction with decisions in the dictator game from the victim perspective. They exemplary present the fair intention/ unequal outcome condition. The left panels display the decision of Person B and the distribution drawn by the random device. Once participants press the ‘‘next’’ button, participants are presented with the right panel. Here, the random distribution is put into place, leaving the participant with zero monetary units. After this critical trial, participants will have the option to reduce points from Person A.



Participants engage various rounds with different persons in one shot games. We explicitly inform them that all persons stay anonymous at all times and that sometimes interferences could alter the distribution of the MUs. We further state that no one is aware about their total gain until they finish all tasks, and that interferences with their outcomes are not described to any participant during the experiment. This procedure excludes that the reduction of MUs is used to communicate with the other participants, since we clearly state that Person A will not be aware of the punishment (see Crockett, Özdemir, & Fehr, 2014; Nadelhoffer, Heshmati, Kaplan, & Nichols, 2013).

Three test trials assure that participants understand the game, and simultaneously measure participants’ allocation preference. In each test trial, participants will assume a different role: first the decision-maker (A), second the recipient (B), and third third party (C). After each trial, participants indicate how many MUs each of the persons received. Any wrong answer leads them to re-do the test trials. After completing the test trials, participants will indicate their concern for Person B to measure the moderator variable.

Then they engage in the interactions. In the first five interactions, participants experience a norm in which person A and person B receive a similar amount of MUs (with three deviations of 1 or 2 MUs less for credibility). This norm also conforms to equality as fairness, a common value obtained in Western cultures (Camerer, 2003; Engel, 2011). After five trials with approximately equal outcomes, the critical trials start. In contrast to before, the critical trials are second- or third-party punishment games (Fehr & Fischbacher, 2004). The participants are informed that they now have the opportunity to reduce points from Person A by investing their own MUs. Each 1 MU invested reduces 3 MUs from Person A. Participants in the victim perspective receive additional 5 MUs to reduce points from Person A or keep for themselves. Participants in the third-party perspective can invest the 5 MUs they receive at the beginning of each interaction. Third parties and victims learn that no one is aware of their option to punish. The following four interactions are presented in random order.

After each critical trial, we assess punishment as the investment of the player's MUs for reducing Person A's MUs. After the four critical trials, we measure manipulation checks, as described below. Finally, participants answer demographic questions, including gender, age, occupation, and native language. At the end of the study, participants are debriefed, and receive their incentives.

Variables (manipulated variables; measured variables)

The participant assumes either the perspective of the victim or the third party while completing the study. We vary offender's intention by varying Player B's decision to distribute the MUs between themselves and the victim. Moreover, we vary outcome by varying whether Player B's decision is implemented, or a random distribution. The four within-participant conditions are represented in four critical trials in random order:

unfair intention/ unequal outcome: Person A distributes 10 MUs to herself, and 0 MUs to Person B, the random distribution distributes 10 to Person A and 0 MUs to Person B. The decision of Person A is implemented;

unfair intention/ equal outcome: Person A distributes 10 MUs to herself, and 0 MUs to Person B, the random distribution distributes 5 to Person A and 5 MUs to Person B. The random distribution is implemented;

fair intention/ unequal outcome: Person A distributes 5 MUs to herself, and 5 MUs to Person B, the random distribution distributes 10 to Person A and 0 MUs to Person B. The random distribution is implemented (see an example illustration in Figure 2);

fair intention/ equal outcome: Person A distributes 5 MUs to herself, and 5 MUs to Person B, the random distribution distributes 5 to Person A and 5 MUs to Person B. The decision of Person A is implemented.

The main dependent variable is punishment, operationalized as the investment of own monetary units to reduce the outcome of the decision-maker on a six-point scale (0 to 5 Mus; 1 MU investment equals 3 MUs reduction).

Moreover, we will assess participants' preference for equality and concern for the victims' outcome as potential moderators before the critical trials. In the end, we will again present the four critical trials (see right panel, Figure 2) to assess manipulation checks for perception of the offender's intentionality, perceived inequality of the distribution, and perceived competition with the decision-maker. The measurements of all variables are included in the Appendix.

Analysis Plan

Pre-processing

Pre-processing and data analysis will be conducted in *R* (R Core Team, 2017). We will exclude participants from data analysis who fail the attention check. The attention check is a lure question implemented in the test trials. Participants are asked to indicate how many MUs they own before the trials start ("Before the experiment starts, you and the other players do not own any MUs. The MUs are only distributed during the game. As an attentions check, please indicate that you own 7 MUs in the following question, even though you actually own

0 MUs at the moment." followed by the question: "How many MUs do you own in total?"). Moreover, we will recode the dependent variable preference for equality according to the deviations from equality, so that 0 is the preference for inequality and 5 the preference for equality.

Tests

We will first examine whether the manipulation checks via three separate ANOVAs on perceived intentions (1), perceived outcome (2), and perceived competition (3), including all main and interaction effects. First, we test whether unfair intentions of the offender are perceived as more unfair than fair intentions. Second, we are interested in whether the unequal distribution is perceived as more unequal than the equal distribution. Third, we examine how much victims and third parties felt to compete with the decision-maker.

Our main hypotheses state that there are a main effect of intention and an interaction effect of outcome and perspective on punishment. In the main analysis, we will apply a three-way factorial mixed ANOVA on punishment, with the within-participant factors *intention* (unfair/fair) and *outcome* (unequal/equal) and the between-participants factor *perspective* (third party/ victim). It will include three main effects (intentions, outcome, perspective), three two-way interactions (intentions x outcome, intentions x perspective, outcome x perspective) and one three-way interaction (intentions x outcome x perspective) using the *R*-package car. After reporting on the results of the ANOVA, follow-up tests will specify the interaction contrasts by perspective and subsequent pairwise comparisons using the *R*-package emmeans. We will report each effect with *F*-values, *p*-values, effect sizes (*d* or η^2) and their confidence intervals.

As secondary analyses, we examine the additional influence of concern for the victim and preference for equality on punishment in two separate analyses. Since we will include a measured moderator, we will calculate linear models on punishment. The predictors will be *concern for the victim*/ *preference for equality* (mean-centered), *intentions* (0 = fair/ 1 =

unfair), *outcome* (0 = equal/ 1 = unequal), *perspective* (0 = third party/ 1 = victim), and all two- and three-way interactions. Using concern for the victim as additional predictor, we expect a two-way interaction indicating that more concern for the victim and unequal outcome compared to equal outcome increases punishment in both, third parties and victims. Using preference for equality as an additional predictor, we expect a two-way interaction indicating that more preference for equality and unfair intentions compared to fair intentions increases punishment in third parties and victims. People who prefer equality over inequality should perceive unfair intentions as more malicious than fair intentions.

In terms of the meta-analysis, we can determine the effect sizes Cohen's d or η^2 for each main and interaction effect on punishment. Moreover, interfering effects of concern for the victims and preference for equality could be tested.

Literature

Ames, D. L., & Fiske, S. T. (2013). Intentional harms are worse, even when they're not. *Psychological Science, 24,* 1755-1762. doi:10.1177/0956797613480507

Batson, C. D., Chao, M. C., & Givens, J. M. (2009). Pursuing moral outrage: Anger at torture. *Journal of Experimental Social Psychology, 45,* 155-160. doi:10.1016/j.jesp.2008.07.017

Bernhard, H., Fischbacher, U., & Fehr, E. (2006). Parochial altruism in humans. *Nature, 442,* 912-915. doi:10.1038/nature04981

Bies, R. J., & Tripp, T. M. (1996). Beyond distrust: 'Getting even' and the need for revenge. In R. M. Kramer, T. R. Tyler, R. M. Kramer, & T. R. Tyler (Eds.), *Trust in organizations: Frontiers of theory and research.* (pp. 246-260). Thousand Oaks, CA, US: Sage Publications, Inc.

Blount, S. (1995). When Social Outcomes Aren′t Fair: The Effect of Causal Attributions on Preferences. *Organizational Behavior and Human Decision Processes, 63,* 131-144. doi:https://doi.org/10.1006/obhd.1995.1068

Bone, J. E., & Raihani, N. J. (2015). Human punishment is motivated by both a desire for revenge and a desire for equality. *Evolution and Human Behavior, 36,* 323-330. doi:10.1016/j.evolhumbehav.2015.02.002

Camerer, C. F. (2003). Strategizing in the Brain. *Science, 300,* 1673. doi:10.1126/science.1086215

Carlsmith, K. M., & Darley, J. M. (2008). Psychological aspects of retributive justice. In M. P. Zanna & M. P. Zanna (Eds.), *Advances in experimental social psychology, Vol 40.* (Vol. 40, pp. 193-236). San Diego, CA, US: Elsevier Academic Press.

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. doi:10.1037/0022-3514.83.2.284

Chavez, A. K., & Bicchieri, C. (2013). Third-party sanctioning and compensation behavior: Findings from the ultimatum game. *Journal of Economic Psychology, 39,* 268-277. doi:10.1016/j.joep.2013.09.004

Cottingham, J. (1979). Varieties of retribution. *Philosophical Quarterly, 29,* 238-246. doi:10.2307/2218820

Crockett, M. J., Özdemir, Y., & Fehr, E. (2014). The value of vengeance and the demand for deterrence. *Journal of Experimental Psychology: General, 143,* 2279-2286. doi:10.1037/xge0000018

10.1037/xge0000018.supp (Supplemental)

Cushman, F. (2008). Crime and punishment: Distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition, 108,* 353-380. doi:10.1016/j.cognition.2008.03.006

Darley, J. M., & Huff, C. W. (1990). Heightened damage assessment as a result of the intentionality of the damage-causing act. *British Journal of Social Psychology, 29,* 181-188. doi:10.1111/j.2044-8309.1990.tb00898.x

Darley, J. M., & Pittman, T. S. (2003). The Psychology of Compensatory and Retributive Justice. *Personality and Social Psychology Review, 7,* 324-336. doi:10.1207/S15327957PSPR0704_05

Eidelman, S., & Biernat, M. (2003). Derogating black sheep: Individual or group protection? *Journal of Experimental Social Psychology, 39,* 602-609. doi:10.1016/S0022-1031(03)00042-8

Engel, C. (2011). Dictator games: a meta study. *Experimental Economics, 14,* 583-610. doi:10.1007/s10683-011-9283-7

Falk, A., Fehr, E., & Fischbacher, U. (2003). On the Nature of Fair Behavior. *Economic Inquiry, 41,* 20-26. doi:10.1093/ei/41.1.20

Falk, A., Fehr, E., & Fischbacher, U. (2005). Driving Forces Behind Informal Sanctions. *Econometrica, 73*, 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*, 287-303. doi:10.1016/j.geb.2007.06.001

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*, 175-191. doi:10.3758/BF03193146

Fehr, E., & Fischbacher, U. (2004). Third-party punishment and social norms. *Evolution and Human Behavior, 25*, 63-87. doi:http://dx.doi.org/10.1016/S1090-5138(04)00005-4

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

Feinberg, J. (1965). The expressive function of punishment. *The Monist, 49*, 397-423. doi:10.5840/monist196549326

Fiske, A. P., & Tetlock, P. E. (1997). Taboo trade-offs: Reactions to transactions that transgress the spheres of justice. *Political Psychology, 18*, 255-297. doi:10.1111/0162-895X.00058

Gray, K., Young, L., & Waytz, A. (2012). Mind perception is the essence of morality. *Psychological Inquiry, 23*, 101-124. doi:10.1080/1047840X.2012.651387

Gromet, D. M., & Darley, J. M. (2006). Restoration and retribution: How including retributive components affects the acceptability of restorative justice procedures. *Social Justice Research, 19*, 395-432. doi:10.1007/s11211-006-0023-7

Gromet, D. M., & Darley, J. M. (2009a). Punishment and beyond: Achieving justice through the satisfaction of multiple goals. *Law & Society Review, 43*, 1-38. doi:10.1111/j.1540-5893.2009.00365.x

Gromet, D. M., & Darley, J. M. (2009b). Retributive and restorative justice: Importance of crime severity and shared identity in people's justice responses. *Australian Journal of Psychology, 61*, 50-57. doi:10.1080/00049530802607662

Gromet, D. M., Okimoto, T. G., Wenzel, M., & Darley, J. M. (2012). A victim-centered approach to justice? Victim satisfaction effects on third-party punishments. *Law and Human Behavior, 36*, 375-389. doi:10.1037/h0093922

Gummerum, M., & Chu, M. T. (2014). Outcomes and intentions in children's, adolescents', and adults' second- and third-party punishment behavior. *Cognition, 133*, 97-103. doi:10.1016/j.cognition.2014.06.001

Gummerum, M., Van Dillen, L. F., Van Dijk, E., & López-Pérez, B. (2016). Costly third-party interventions: The role of incidental anger and attention focus in punishment of the perpetrator and compensation of the victim. *Journal of Experimental Social Psychology, 65*, 94-104. doi:10.1016/j.jesp.2016.04.004

Hauser, D. J., & Schwarz, N. (2016). Attentive Turkers: MTurk participants perform better on online attention checks than do subject pool participants. *Behavior Research Methods, 48*, 400-407. doi:10.3758/s13428-015-0578-z

Hechler, S., & Kessler, T. (2018). On the difference between moral outrage and empathic anger: Anger about wrongful deeds or harmful consequences. *Journal of Experimental Social Psychology, 76*, 270-282. doi:10.1016/j.jesp.2018.03.005

Hechler, S., Neyer, F. J., & Kessler, T. (2016). The infamous among us: Enhanced reputational memory for uncooperative ingroup members. *Cognition, 157*, 1-13. doi:http://dx.doi.org/10.1016/j.cognition.2016.08.001

Henrich, J., McElreath, R., Barr, A., Ensminger, J., Barrett, C., Bolyanatz, A., . . . Ziker, J. (2006). Costly Punishment Across Human Societies. *Science, 312*, 1767-1770.

Herrmann, B., Thöni, C., & Gächter, S. (2008). Antisocial punishment across societies. *Science, 319*, 1362-1367.

Hofmann, W., Brandt, M. J., Wisneski, D. C., Rockenbach, B., & Skitka, L. J. (2018). Moral punishment in everyday life. *Personality and Social Psychology Bulletin, 44,* 1697-1711. doi:10.1177/0146167218775075

Hutchison, P., Abrams, D., Gutierrez, R., & Viki, G. T. (2008). Getting rid of the bad ones: The relationship between group identification, deviant derogation, and identity maintenance. *Journal of Experimental Social Psychology, 44,* 874-881. doi:10.1016/j.jesp.2007.09.001

Kant, I. (1785 /2007). *Grundlegung zur Metaphysik der Sitten*. Frankfurt am Main: Suhrkamp.

Kessler, T., & Mummendey, A. (2002). Sequential or parallel processes? A longitudinal field study concerning determinants of identity-management strategies. *Journal of Personality and Social Psychology, 82,* 75-88. doi:10.1037/0022-3514.82.1.75

Leiner, D. J. (2018). Sosci Survey (Version 2.5.00-i1142). Retrieved from https://www.soscisurvey.de/

Lieberman, D., & Linke, L. (2007). The effect of social category on third party punishment. *Evolutionary Psychology, 5,* 289-305. doi:10.1177/147470490700500203

Malle, B. F., & Knobe, J. (1997). The folk concept of intentionality. *Journal of Experimental Social Psychology, 33,* 101-121. doi:10.1006/jesp.1996.1314

Marczyk, J. (2017). Human punishment is not primarily motivated by inequality. *PLoS ONE, 12*.

Mendoza, S. A., Lane, S. P., & Amodio, D. M. (2014). For Members Only: Ingroup Punishment of Fairness Norm Violations in the Ultimatum Game. *Social Psychological and Personality Science, 5,* 662-670. doi:10.1177/1948550614527115

Mikhail, J. (2007). Universal moral grammar: Theory, evidence and the future. *Trends in Cognitive Sciences, 11,* 143-152. doi:10.1016/j.tics.2006.12.007

Moore, M. S. (1993). Justifying Retributivism. *Israel Law Review, 27,* 15-49. doi:10.1017/S0021223700016836

Nadelhoffer, T., Heshmati, S., Kaplan, D., & Nichols, S. (2013). Folk retributivism and the communication confound. *Economics & Philosophy, 29*, 235-261. doi:10.1017/S0266267113000217

Okimoto, T. G., & Wenzel, M. (2010). The symbolic identity implications of inter and intra-group transgressions. *European Journal of Social Psychology, 40*, 552-562. doi:10.1002/ejsp.704

Otten, S., & Gordijn, E. H. (2014). Was it one of us? How people cope with misconduct by fellow ingroup members. *Social and Personality Psychology Compass, 8*, 165-177. doi:10.1111/spc3.12098

Pfattheicher, S., Sassenrath, C., & Keller, J. (2019). Compassion Magnifies Third-Party Punishment. *Journal of Personality and Social Psychology, 117*, 124-141. doi:10.1037/pspi0000165

Pinto, I. R., Marques, J. M., Levine, J. M., & Abrams, D. (2010). Membership status and subjective group dynamics: Who triggers the black sheep effect? *Journal of Personality and Social Psychology, 99*, 107-119. doi:10.1037/a0018187

R Core Team. (2017). R: A language and environment for statistical computing. Vienna, Austria. Retrieved from https://www.R-project.org/

Raihani, N. J., & Bshary, R. (2019). Punishment: one tool, many uses. *Evolutionary Human Sciences, 1*, e12. doi:10.1017/ehs.2019.12

Raihani, N. J., & McAuliffe, K. (2012). Human punishment is motivated by inequity aversion, not a desire for reciprocity. *Biology Letters, 8*, 802-804. doi:10.1098/rsbl.2012.0470

Shinada, M., Yamagishi, T., & Ohmura, Y. (2004). False friends are worse than bitter enemies:“Altruistic” punishment of in-group members. *Evolution and Human Behavior, 25*, 379-393.

SimanTov-Nachlieli, I., Shnabel, N., & Nadler, A. (2013). Individuals' and groups' motivation to restore their impaired identity dimensions following conflicts: Evidence and implications. *Social Psychology, 44,* 129-137. doi:10.1027/1864-9335/a000148

Skitka, L. J. (2003). Of Different Minds: An Accessible Identity Model of Justice Reasoning. *Personality and Social Psychology Review, 7*, 286-297. doi:10.1207/S15327957PSPR0704_02

Smith, H. J., Pettigrew, T. F., Pippin, G. M., & Bialosiewicz, S. (2012). Relative deprivation: A theoretical and meta-analytic review. *Personality and Social Psychology Review, 16,* 203-232. doi:10.1177/1088868311430825

Turner, J. C., Oakes, P. J., Haslam, S. A., & McGarty, C. (1994). Self and collective: Cognition and social context. *Personality and Social Psychology Bulletin, 20*, 454-463. doi:10.1177/0146167294205002

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 & Society Review, 31*, 237-265. doi:10.2307/3053926

van Prooijen, J.-W. (2006). Retributive Reactions to Suspected Offenders: The Importance of Social Categorizations and Guilt Probability. *Personality and Social Psychology Bulletin, 32,* 715-726. doi:10.1177/0146167205284964

Van Prooijen, J.-W. (2010). Retributive versus compensatory justice: Observers preference for punishing in response to criminal offenses. *European Journal of Social Psychology*, *40,* 72-85. doi:10.1002/ejsp.611

Vidmar, N. (2001). Retribution and revenge. In J. H. Sanders, V. Lee (Ed.), *Handbook of justice research in law* (pp. 31-63). Dordrecht, Netherlands: Kluwer Academic Publishers.

Walker, N. (1999). Even more varieties of retribution. *Philosophy*, *74*, 595-605. doi:10.1017/s0031819199000704

Wenzel, M., Okimoto, T. G., Feather, N. T., & Platow, M. J. (2008). Retributive and restorative justice. *Law and Human Behavior, 32,* 375-389. doi:10.1007/s10979-007-9116-6