

**From the holdings of the Department of Special Collections , Princeton University Library,
Princeton, NJ**

These documents are either in the public domain, or can only be used for educational and research purposes (Fair use) as per U.S. Copyright law (text below). By accessing this file, all users agree that their use falls in the public domain or within fair use as defined by the copyright law. We do not charge any permission or usage fees for the publication of images of material in our collections, including those provided by Princeton University Library, via our website, catalog, or directly from staff. For more information regarding copyright, credit, and citation please review the information on the Copyright, Credit and Citation Guidelines page of our website.

<https://library.princeton.edu/special-collections/policies/copyright-credit-and-citation-guidelines>. If you have any questions, including how to order higher-quality images from the collection, contact Special Collections Public Services via our Ask Us! Form. <https://library.princeton.edu/special-collections/ask-us>

Doctoral Dissertations, 2011-2026

Psychology

2015

Beyond retribution: The role of transformative justice motives for people's reactions to wrongdoers

Funk, Friederike

Princeton, NJ : Princeton University

<http://arks.princeton.edu/ark:/88435/dsp017s75df69c>

U.S. Copyright law: The copyright law of the United States (Title 17, §108, United States Code) governs the making of photocopies or other reproductions of copyrighted material. Under certain conditions specified in the law, libraries and archives are authorized to furnish a photocopy or other reproduction. One of these specified conditions is that the photocopy or reproduction is not to be used for any purpose other than private study, scholarship, or research. If a user makes a request for, or later uses, a photocopy or reproduction for purposes in excess of fair use that user may be liable for copyright infringement. This institution reserves the right to refuse to accept a copying order if, in its judgment, fulfillment of the order would involve a violation of copyright law. <https://www.copyright.gov/title17/title17.pdf>

BEYOND RETRIBUTION:
THE ROLE OF TRANSFORMATIVE JUSTICE MOTIVES
FOR PEOPLE'S REACTIONS TO WRONGDOERS

Friederike Funk

A DISSERTATION
PRESENTED TO THE FACULTY
OF PRINCETON UNIVERSITY
IN CANDIDACY FOR THE DEGREE
OF DOCTOR OF PHILOSOPHY

RECOMMENDED FOR ACCEPTANCE
BY THE DEPARTMENT OF PSYCHOLOGY

Advisor: Deborah A. Prentice

June 2015

© Copyright by Friederike Funk, 2015 All rights reserved.

Abstract

The present dissertation focuses on the role of transformative justice motives for people's reactions to wrongdoing. Findings from Studies 1a and 1b revealed punishment to be satisfying for punishers only if transgressors relate punishment with a change in attitude. Moreover, findings from Study 2 demonstrated that the effect of transgressor change on justice-related satisfaction is different from mere deterrence and not reducible to retributive effects of inferred suffering. Study 3 examined the transgressor's display of remorse as one important marker of authenticity of change. Finally, Study 4 explored the origins of transformative justice motives and found they are more relevant in ingroup than in outgroup contexts. While Studies 1 and 2 only looked at people who decided to punish, Studies 3 and 4 also included people who decided not to punish. Findings suggest that transformative motives seem to generally arise after transgression and are expressed through punishment or other means. All in all, the present set of studies highlights that people do not only seem to give transgressors what they deserve. They hope to transform transgressors and make them understand that what they did was wrong. As social beings, people seem to strive for reconciliation and hope to achieve a change in attitude in those who have treated them unfairly.

Table of Contents

Abstract.....	i.....
Acknowledgments.....	iv
Introduction.....	1
Transformative justice motives.....	6
Transformative justice motives apply to punishers and ~~punishers~~	9
Affective consequences of transgressor change.....	9
Overview of studies.....	12
Study 1a: Initial evidence for transformative justice motives.....	12
Method.....	13
Results.....	16
Discussion.....	19
Replication Study 1b: Transgressor change affects ~~related~~ satisfaction...	20
Method.....	21
Results.....	22
Discussion.....	24
Study 2: Transformative justice motives are distinguishable from	
deterrent and retributive motives.....	25
Methods.....	30
Results.....	34
Discussion.....	39

Study 3: Transformative justice motives aim at authentic change	
on affective, behavioral, and cognitive levels.....	43
Creation and validation of facial stimuli.....	45
Method.....	48
Results.....	53
Discussion.....	59
Study 4: The social origins of transformative justice motives.....	62
Method.....	64
Results.....	69
Discussion.....	78
General Discussion.....	85
Implications for the legal system.....	91
Next steps for psychological research on transformative justice motives	
.....	93
Conclusion.....	97
References.....	99
Appendix	116

Acknowledgments

This research was supported by a Fellowship of Woodrow Wilson Scholars (2012-14) and a Graduate Prize Fellowship by Princeton University Center for Human Values (2014-15).

Studies 1a and 1b have been published as Studies 2a and 2b in Funk, F., McGeer, V., & Gollwitzer, M. (2014). Get the message: Punishment is satisfying if the transgressor responds to its communicative intent. *Personality and Social Psychology Bulletin*, 40(8), 986-997.

Thanks go to Lauren Feldman for providing invaluable research assistance in Studies 2 and 4, as well as to Mirella Walker for creating and sharing the facial stimuli for Study 3.

Beyond retribution:

The role of transformative justice motives for people's reactions to wrongdoers

Transgressions and punishment are everywhere in our lives. We read about them in the newspapers and in books. We follow court trials on TV and spend much of our leisure time watching fictional TV shows that are about bringing justice to a person who committed a crime. In our own lives, we have a desire to punish people who treat us unfairly: we honk at cars that cut us off the road, for instance, and we give friends who forgot our birthday the silent treatment.

People's desire to restore justice is strong and can often be in contrast to their self interest (for a review see Lerner & Clayton, 2011) and people punish even when it is costly to do so (Fehr & Gächter, 2002; Herrich et al., 2005). Research findings suggest that the desire to punish is not linked to cathartic motives (i.e., acting out on aggression would be, see Bushman, Baumeister, & Phillips, 2001). Instead, people even punish when they know that punishment will not improve their mood when they have taken a bogus mood-freezing pill, see Gollwitzer & Bushman, 2012. But what is it that drives people to punish?

Psychologists have broadly offered two possible responses to this question, drawing from the philosophical distinction between deontological justifications for punishment (beginning with Kant, 1797/1996) and consequentialist justifications (beginning with Bentham, 1830/2002). See e.g., Carlsmith, 2006; Darley, Carlsmith, & Robinson, 2000; Keller, Oswald, Stucki, & Gollwitzer, 2010. Deontological justifications refer to retributive, backward-looking concerns (giving offenders their •just

deserts, which is the proportional punishment they deserve in light of the wrong they have done), whereas consequentialist justifications are mostly concerned with utilitarian, forward-looking reasons for punishment (e.g., the prevention of future crimes).

When people are asked why offenders should generally be punished, they appear to endorse both retributive and consequentialist justifications (Carlsmith, 2008; Darley, 2002). However, empirical findings have revealed that when people are asked to assign punishment to specific transgressions, they are affected more by the seriousness of the offense than by its likelihood of detection or offense frequency (Carlsmith, Darley, & Robinson, 2002) and by the likelihood that the transgressor can commit future offenses (Darley et al., 2000). People punish proportionally to a transgression's seriousness and recommend punishment even when there is no deterrent or incapacitating value to it. These reactions occur intuitively and heuristically (Aroni & Fridlund, 2012; Carlsmith, 2006; Darley, 2009). Psychologists have therefore concluded that people's punishment behavior is influenced by retributive motives and not consequentialist motives (for a review see Carlsmith & Darley, 2008).

Deterrent motives, however, do not exhaust the complete range of possible consequentialist goals. There are two different kinds of specific consequentialist punishment goals that both ultimately aim at behavior control to prevent a transgressor from reoffending: negative prevention (deterrence) and positive prevention (also referred to as confirmation of norms, offender reform, rehabilitation, or reeducation; Kellee et al., 2010; Orth, 2003; Oswald, Hupfeld, Klug, & Gabriel, 2002; Widmar & Miller, 1980). Negative prevention refers to the use of punishment as a sanctioning threat to deter transgressors from reoffending, i.e., these transgressions do not pay. This punishment

goal has been extensively studied and repeatedly ruled out as main motivation for punishment behavior (see above). Positive prevention, on the other hand, refers to creating a sense about what is right and wrong in the transgressor. This punishment goal has not received the same attention in experimental research designs. As a consequence, because research findings have repeatedly ruled out negative prevention as concern that motivates punishment behavior, it is not valid to conclude that retributive motives are the dominant punishment motives (see, e.g., Carlsmith & Darley, 2008; Crockett, Özdemir, & Fehr, 2014; Keller et al., 2010). The present dissertation aims to study if people punish for reasons of positive prevention.

That punishment might aim at positive prevention is reflected in the way laypeople talk about punishing transgressors. The expression that it teaches a lesson, for instance, refers to educative functions of punishment. The ideal that offenders change as a reaction to punishment is integrated in the American culture of redemptive narratives (Cullen, Pealer, Fisher, Applegate, & Santana, 2008). Offenders can be saved or corrected, and prisons are called correctional institutions. Originally prisons were even called penitentiaries (Fothman, 1971), thus for the purpose of penance, to transform offenders into law-abiding people. Even when ordinary people tell their life stories, they often refer to the experience of suffering as a defining point of redemption that has given them a second chance in life (McAdams, 2013).

Scholars have also linked the goal of positive prevention to the sphere of punishment. Adam Smith (1759/1869) already noted that „[t]o bring him back to a more just sense [...] of the wrong that he has done to us, is frequently the principal end posed in our revenge, which is always imperfect when it cannot attain this“ (p. 139).

Similarly, the psychologist Fritz Heider (1958) theorized many years ago that resentment is a wish to produce a change in the underlying belief-attitude structure of the attacker, and revenge is the means of realizing this wish (p. 267). For (2004), educating the offender (along with restoring self-esteem) is one of the most powerful goals that people have when they are treated disrespectfully and feel a sense of injustice.

Psychologists have referred to the sense that arises in transgressors who understand that what they did was wrong as a sense of guilt (Baumeister, Stillwell, & Heatherton, 1994) and studied it for instance with expressions of remorse. Research on punishment severity suggests that people have less a desire to punish transgressors who show remorse, indicated by less severe punishment recommendations (Diony & Schlenker, 1989; Eisenberg, Garvey, & Wells, 1998; Gold & Weiner, 2000; Robinson, Smith-Lovin, & Tsoudis, 1994; Rumsey, 1976; Tsoudis & Smith-Lovin, 1998). These findings are in line with positive prevention as punishment motive. If change in the transgressor has already partly occurred, as indicated by the presence of remorse or a sense of guilt, people's need to punish is reduced.

Assuming that punishment serves several goals, the present dissertation examines positive prevention as punishment goal. The used approach purposefully departs from previous psychological traditions that have contrasted retributive punishment goals on the one hand with consequentialist punishment goals on the other hand. Punishment is hypothesized to be retributive (i.e. proportional) in its form but transformative in its scope. As A.C. Ewing (1927, p. 305): 'The retributive theory insists that punishment ought to be for a past offence, utilitarianism that it ought to be for a future good, but may it not be for both? Perhaps it

can only play the latter role effectively by first playing the former, and perhaps playing the former role it ipso facto plays the latter

Philosophical theorists have long recognized that punishment is probably not either purely retributive or purely consequentialist. But some philosophers, punishment is conceptualized as communication of condemnation (see e.g. Feinberg, 1965; Nozick, 1981). Punishment as an infliction of pain expresses disapproval, and greater inflicted pain is a sign of greater disapproval (Ewing, 1929). Although retrospective to a certain degree, condemnation or censure can also be used as an attempt to make offenders understand that their behavior was wrong and to make them change (see e.g. Duff, 2001; Hampton, 1984, 1992; Morris, 1981). Punishment educates transgressors about the wrongness of their actions and thereby produces a transformation in their attitudes is proportional to the wrong committed not to give transgressors their just desert but to show them how wrong their behavior was and how much they need to change

Similarly, atonement theories suggest that positive prevention might be the scope of punishment. Proportional punishment is used as communication and as a mechanism for moral reform so that offenders realize the wrong that they committed (Garvey, 1998, p. 740). Punishment aims at making transgressors realize the wrong committed, which ideally results in a change of their attitude and behavior (Garvey, 1998, 1999). This change is called atonement (Garvey, 1999) or secular penance (Duff, 2001). Such an account of punishment as atonement can be found in many religions and has shaped criminal justice policies in many countries (Sorringe, 1996).

Following philosophers and laypeople's conceptualization of punishment as aiming at a change in the transgressor, the present dissertation looks at the psychological

reality of such a punishment goal. No people punish for reasons of positive prevention, that is do they aim at effecting a change in the offender's attitude. The change in attitude is conceptualized as the attitude toward the transgressor's previous wrongdoing: By changing their attitude, transgressors realize that their behavior was wrong, they feel bad or remorseful about that wrong, and as a result change their behavior by not doing the wrong again. If people punish for reasons of positive prevention, punishment might ultimately aim at reconciliation between the transgressor and the victim and at a restoration of group functionality.

Transformative justice motives

Throughout the present dissertation, the term ,transformative justice motives' is used to refer to people's desire to effect a change in the transgressor's attitude.

Transformation in the transgressor's attitude consists of affective, behavioral, and cognitive components. As mentioned before, on the affective level, the transgressor is supposed to feel remorseful about the wrongful action. On the behavioral level, the transgressor is supposed to refrain from committing the wrongful action again. And on the cognitive level, the transgressor understands that the committed action was wrong.

The transformative justice motive proposed in this dissertation is distinct from but related to notions of moral education, rehabilitation, specific deterrence, and restorat

Transformation versus moral education. The term moral education suggests that transgressors need to be educated about right behaviors and might not have known that what they did was wrong. It is more realistic for most contexts that transgressors technically knew that their behavior would harm another person but that they did not care about it. Punishment is supposed to draw transgressors' attention to the wrong

committed and thereby effect a change in their attitude to it. The term moral education would be too limited as it refers to cognitive knowledge about things, whereas a change in attitude also goes along with a change in affect and behavior. Offenders need to confront and to respond to what they did (see Duff, 2001, p. 98). Although reeducation is ideally the result of punishment, the underlying motive of effecting a visible transformation in the transgressor's attitude is probably more similar to persuasion than to education (see also Duff, 2001, p. 80).

Transformation versus rehabilitation. As mentioned above, positive prevention is also sometimes called offender reform or rehabilitation. Thus, transformation and rehabilitation are very similar to each other. In the criminological literature, however, the term rehabilitation has often been used to refer to the form of punishment (e.g., rehabilitative punishment such as sentencing offenders to therapy) or to refer to add-on practices to punishment (e.g., offenders undergo rehabilitative practices when they are in prison) that help offenders to return to an acceptable way of life and to reintegrate into society (Cullen et al., 2002). The present dissertation uses the term transformation to be precise about the scope of the research, which is studying the psychological motive of a wronged person and not the practice of punishment.

Transformation versus specific deterrence. At first sight, it seems that effecting a change in the offender's attitude is the punishment goal as specific deterrence because both goals are forward-oriented and ideally result in a lower risk of reoffending. Yet, deterrence (negative prevention) refers to a cost-benefit analysis about possible transgressions and potentially resulting punishments, leading transgressors to conclude that the committed behavior does not pay. Positive prevention refers to a

genuine change in the offender's attitude that goes beyond mere behavior change. The offender understands that certain behaviors are wrong and, as a consequence, will not reoffend. As mentioned above, transformation consists of a genuine change in attitude ideally consisting of affective, behavioral, and cognitive components. Feeling remorse and understanding that what one did was wrong would be irrelevant for reasons of pure deterrence as long as the transgressor demonstrates a change in behavior (i.e. is deterred and does not reoffend).

Transformation versus restoration. Norm violations question the validity of norms and values (Durkheim, 1893/1964; Vidmar, 2001). Although the term, restorative justice, is used in different ways by different scholars (for instance sometimes including retributive notions and sometimes not, for a review see Daly & Pfeiffer, 2011), the general goal of restorative justice procedures is to reaffirm a shared value consensus in a bilateral process (Wenzel, Okimoto, Feather, & Platow, 2008, p. 375). In contrast to a unilateral imposition of punishment, restorative justice procedures include all involved parties (i.e. the victim, the offender, and the community) and try to find ways to repair the harm that has been caused by the offense (Bazemore, 1998; Braithwaite, 1998). Restorative justice settings try to heal the wound that an offense has caused (Zehr, 1990). The term restoration suggests that a previous state can be restored. In the context of restorative justice settings, this restoration refers to a consensus about norms and values. Because the scope of the present dissertation is to study psychological motives and victimized people have not necessarily had value consensus with transgressors prior to a transgression, transformation is a more precise term. Generally, it refers to the goal of effecting a different, new state. Moreover, this motive primarily

refers to a transformation in the transgressor and is not primarily directed at the healing of wounds of the victim or the community. Lastly, the use of a different term accounts for the unilateral nature of transformative justice motives in contrast to the bilateral processes in restorative justice settings.

Transformative justice motives apply to punishers and nonpunishers

As the term suggests, previous work on punishment motives has only focused on people who punish. Punishment is usually understood as pain or loss that is intentionally and openly inflicted on [a person] in retaliation for something [that person] did (Deigh, 2014, p. 190). Many people who are treated unfairly, however, decide not to punish transgressors but react in different, potentially even constructive ways (McCullough, 2008). Traditionally, retributive accounts on punishment give little attention to people who refrain from punishing. In order to fully understand the motives that arise in the aftermath of a transgression, however, it is important that a psychological theory on punishment motives can also explain why some people decide not to give transgressors what they deserve. The present dissertation aims at developing a general account on transformative justice motives (i.e. positive prevention) that refers to punishers and nonpunishers alike. People who have been treated unfairly are hypothesized to aim at a transformation in the transgressor's attitude, even when they refrain from punishing and react with different behavior, rendering transformative justice motives a powerful aspect of people's reactions to wrongdoing.

Affective consequences of transgressions change

In order to study people's transformative justice motives, the present dissertation examines when punishment (or no punishment) leaves people satisfied. Brain imaging

studies suggest that people anticipate satisfaction when they make decisions about punishing a transgressor (de Quervain et al., 2004). Because anticipated satisfaction from punishment is not the reason why people punish (Gollwitzer & Bushman, 2012), it is interesting to examine both what people think will make punishment satisfying and what actually does make it satisfying in order to identify people's underlying punishment motives.

Previous studies have suggested that there is supposedly a match between the satisfaction people anticipate from punishing transgressors and the satisfaction they actually experience when they punish transgressors. In these studies, participants either imagined punishing and predicted how satisfied punishment would leave them, or they actually punished and indicated how satisfied punishment actually left them. Participants in these studies anticipated punishment to be satisfying but did not experience punishment that way. Instead, actual punishers were left dissatisfied, leading the researchers to conclude that punishment has paradoxical consequences (Crosby, Wilson, & Gilbert, 2008).

More recent studies have partly explained these effects by investigating the role of feedback from transgressors for punishers' satisfaction from punishment. Several studies have found that punishers are only satisfied after punishing if they receive feedback from the punished person acknowledging the punishment (Gollwitzer & Denzler, 2009; Gollwitzer, Meder, & Schmitt, 2011). More precisely, punishers were only satisfied after punishment if transgressors let punishers know that they understand the bad treatment they have received to be a reaction to their previous behavior. If punishers did not receive any feedback from transgressors, however, punishment did not

lead to satisfaction, replicating what Carlsmith and colleagues (2008) had found in their studies.

To add to this line of research, a recent study examined the question of whether punishers are accurate in predicting the difference that transgressor feedback makes (Funk, McGeer, & Gollwitzer, 2014, Study 1). In this study, participants were asked to either imagine receiving feedback after punishment (which is similar to Gollwitzer and colleagues' previous studies, but this time examining the anticipation of punishment) or to imagine receiving no feedback after punishment (which is similar to Carlsmith and colleagues' setup, but this time making the absence of transgressor feedback explicit to participants). Findings revealed that participants only anticipated satisfaction if they imagined getting feedback from the transgressor after punishing, but not if they imagined getting no feedback. Moreover, the anticipated levels of satisfaction did not differ from the experienced levels of satisfaction that actual punishers reported after punishing transgressors who did or did not provide feedback indicating understanding of why they were being punished. Thus, these findings debunked the paradoxical consequences found earlier by explaining them in the context of transgressor feedback.

All in all these findings suggest that satisfaction from punishments dependent on what happens after punishment and that punishment is not satisfying or dissatisfying per se. Therefore, examining actual or imagined satisfaction after punishment is the ideal way to study what punishers might have aimed at when they decided to punish in the first place.

Overview of studies

The four studies presented in this dissertation examined the desire to effect a change in the transgressor by looking at affective consequences if the desire has been met (i.e. the transgressor has changed) or not (i.e. the transgressor has not changed). Studies 1a and 1b examined if change in attitude in the transgressor after punishment leads to more justice-related satisfaction and closure than no change in attitude. Study 2 further aimed at disentangling this effect from deterrence (here behavior change) and inferred increase in suffering that would ultimately support a retributive notion of punishment. Studies 3 and 4 studied people who decided not to punish to better understand the nature of transformative justice motives. Study 3 manipulated the transgressor's display of remorse that went along with the feedback about change in order to examine the importance of authenticity. Finally, Study 4 examined the social origins of transformative justice motives by manipulating the transgressor's social category.

Study 1a: Initial evidence for transformative justice motives

Study 1 investigated satisfaction after actual punishment. If punishers merely aimed at giving people their just deserts and punished for purely retributive reasons, punishment should be satisfying regardless of its consequences (see, e.g., Carlsmith & Darley, 2008, for a review on the psychology of retribution). If punishers used punishment primarily to communicate disapproval and only wanted transgressors acknowledge that the bad treatment they receive is punishment for a previous act by receiving feedback that the transgressor understands what he or she is being punished for, see Gollwitzer et al., 2011, punishment should again be satisfying regardless of its

consequences as long as such feedback is provided¹, however, people have transformative justice motives and use punishment as a tool to transform transgressor as suggested in this dissertation, punishers should only be satisfied after punishment if it has this desired effect on the transgressor.

HYPOTHESIS1:

Punishment is not satisfying per se. It is only satisfying for punishers if it effects a change in the transgressor

Method

Study 1a was a laboratory study that experimentally varied the presence and the type of feedback transgressors would give punishers after punishment and measured punishers' satisfaction. In total, there were four experimental conditions manipulated between subject punishment and feedback about change punishment and feedback about no change punishment without feedback and no punishment no feedback (control condition)

Procedure. Participants (404 undergraduate students, 258 female, $M_{age} = 19.5$, $SD_{age} = 1.1$) were recruited through the University's psychology subject pool for the study, Problem solving in virtual groups¹. They participated in groups of two and were told they would be participating in a study that examines whether virtual collaboration can enhance group performance. Participants were placed in two adjacent rooms and led to believe their computers would be connected with the other students' computers to work collectively on a task. After reading general instructions about the task, participants were invited to send a message to their partner if they wished. To increase

¹ For each of the studies, if the reported subjects are unique subjects that could not participate in any other study presented in this dissertation.

credibility, all of the participants received a message from their ostensible partner
turn (, Hi €fellow student•... What am I supposed to say, just saw that the more
problems we solve the more bonus we get ?Not bad... Garf) on!

Before the study started, participants were asked to indicate how they feel
now, using general Positive Affect Negative Affect (PANA) items (pleased, positive,
satisfied, happy, angry) (negative $\emptyset$), upset $\emptyset$, and irritated $\emptyset$ (... = .90; order
randomized, 1 =not at all, 7 =very much).

For the main task, participants were asked to individually solve as many
anagrams as possible within two minutes. For each anagram solved, the team rece
\$0.10. After the task, participants read how many anagrams they themselves and their
partners had solved correctly (ostensible partners always solved one anagram less than
the participant) and how much money they therefore earned as a team. It was explained
to participants that they and their partners now have to decide how to split the bonus
they have earned. They would both be asked how to split the bonus, and the final split
would be made by averaging both of their answers. In this reward sharing stage of the
procedure after participants had made a choice about the split, participants always saw
that their partner recommended keeping the entire bonus for themselves. Consequently,
the final averaged payoff was very unfair.

At this point, participants in the three punishment conditions received an
apparent error message from the system indicating that they received significantly less
money than their partner and that this happens when a partner suggests splitting up the
bonus in a way that does not reflect their relative performances. Participants were told
that they could reduce their partner's payoff if they wanted to and that this would not

affect their own bonus. Only participants who decided to reduce their partner's bonus received the second experimental manipulation (i.e. different kinds of feedback from the offender). Participants in the control condition (no punishment no feedback) did not have the opportunity to reduce their partner's bonus and did not receive any feedback.

After participants had made the decision to punish, they were randomly assigned to one of three feedback conditions. They either did not receive any feedback about their partner's punishment (punishment without feedback condition), they received a message indicating a change in attitude (Hey you reduced my bonus! Ok was greedy and I now see what's wrong with that ... I shouldn't be such a jerk in situations like this, punishment and feedback about change condition), or they received a message indicating no change in attitude (Hey, you reduced my bonus! Ok was greedy but I don't see what's wrong with that... In situations like this I always try to get as much as I can, punishment and feedback about no change condition).

On the next page, participants who had received feedback were asked to imagine re-doing the task and deciding how to split up the bonus. They were then asked to indicate if they would allocate the bonus the same way again and were told that they would see how their partner replied to this question on the subsequent page. On that page, participants in the punishment and feedback about change condition saw that their ostensible partner had indicated No, I would behave differently in the future, while participants in the punishment and feedback about no change condition saw that their ostensible partner had indicated Yes, I would behave the same way in the future.

Participants then filled out the dependent measures. First, they were asked once again how they feel right now, using the same PANA items as before (... = .87). Next,

participants filled out three items about their satisfaction regarding the interaction with the other player (I feel satisfied, I feel mad, I feel disappointed); $\alpha = .65$, order randomized (1 = not at all, 7 = very much) and answered some filler items on the anagram task and their partner.

If participants had decided to punish, they further indicated to what extent they thought adjusting their partner's bonus made him/her think about his/her behavior, and whether adjusting their partner's bonus had caused a change in his/her attitude (order randomized, all items 1 = not at all, 7 = very much). After filling out some demographic information, participants were debriefed about the true purpose of the study. They were asked if they had been suspicious that the other partner was not real or if they had known about the real purpose of the study before. Participants were thanked and received half an hour of course credit plus \$0.10 for each anagram they solved.

Results

Preliminary analyses. Participants who split the reward extremely unevenly (suggesting 10% or less for themselves and 90% or more for their partner, or vice versa, $n = 11$) were excluded from the analyses as they did not have the same experience of unfair treatment. One participant did not give informed consent at the end of the study; these data were discarded. Twenty-nine participants were in the control condition (no punishment no feedback). The remaining 365 participants were given the opportunity to punish; only 117 of them decided to reduce their partner's bonus (32%) from the students who decided not to punish ($n = 248$) were not used for the analyses.

Excluding nonpunishers from the dataset did not change the characteristics of the sample. There were no differences in age ($t(362) < 1, p = .881$, participants' gender,

$\chi^2(1, N = 365) = 0.50, p = .481$, suspiciousness, $\chi^2(1, N = 365) = 1.00, p = .344$, or differences in the number of solved anagrams, $t(363) < 1, p = .482$, between participants who decided to punish and participants who decided not to punish.

Many students were suspicious that the other partner was not real (e.g., because they knew the student they were paired with, or they had heard from other participants what the study was really about). Those students were excluded from the analyses (n = 47). For the punishment conditions there were more suspicious students in the punishment and feedback about change condition (n = 23) than in the punishment and feedback about no change condition (n = 12), the punishment without feedback condition (n = 8), and the no punishment no feedback condition (n = 4), $\chi^2(3, N = 144) = 10.87, p = .012$. Nevertheless, even if all of the suspicious students were included in the final analyses, the pattern of significant findings would look identical to the results reported below.

The final sample consisted of 97 students, with 23 to 26 participants in each experimental condition. $M_{\text{age}} = 19.5, SD_{\text{age}} = 1.2$, 63 [65%] female participants were equally distributed across conditions. Participants solved on average 8.8 anagrams ($SD = 2.47$). They proposed keeping on average 53% ($SD = 4\%$, min = 50%, max = 60%) of the bonus for themselves. People who decided to punish deducted on average ($SD = \$0.26$) from their partner's bonus.

The number of anagrams solved, participants' initial allocation recommendations, and the magnitude of punishment did not reliably differ between the

² All of the reported p-values in this dissertation are two-sided unless otherwise noted.

experimental conditions ($p = .03875$) and will not be discussed further. These variables also did not correlate with satisfaction after punishment.

Manipulation checks. As expected, participants in the punishment and feedback about change condition agreed more strongly than participants in the punishment and feedback about no change condition that adjusting their partner's bonus made the partner think about his/her behavior change ($M = 6.08$, $SD = 0.93$; no change $M = 4.08$, $SD = 1.79$; Welch's $t(34.5) = 4.86$, $p < .001$, $d = 1.40$) and agreed more strongly that adjusting the bonus has caused a change in their partner's attitude ($M = 4.96$, $SD = 1.40$; no change $M = 2.58$, $SD = 1.69$; $t(46) = 5.30$, $p < .001$, $d = 1.53$).³

Effects of transgressor change on general affect and satisfaction regarding the interaction.

Change in general affect. Participants felt less positive at time 2 (right before they filled out the measure about satisfaction regarding the interaction, $M = 3.98$, $SD = 1.12$) than at time 1 (at the beginning of the study, $M = 5.21$, $SD = 1.06$), $PANA F(1, 93) = 110.90$, $p < .001$, $\eta^2 = .544$. Experimental condition did not significantly affect general affect: condition $F(3, 93) = 1.77$, $p = .159$. Negative changes in general affect were slightly stronger in the no punishment/no feedback group than in the three punishment groups, as indicated by a marginally significant interaction, $PANA \times \text{condition } F(3, 93) = 2.42$, $p = .071$.

Satisfaction regarding the interaction. As predicted, there was a significant effect of the experimental conditions on participants' satisfaction regarding the

³ Whenever two means are compared, Cohen's d was calculated by using the following formula:

$$d = \frac{M_1 - M_2}{\sqrt{((s_1^2 + s_2^2) / 2)}}$$

interaction with the other play, $F(3, 92) = 329, p = .024, \eta^2 = .097$. Planned contrast analyses revealed that participants in the punishment and feedback about change condition ($M = 4.46, SD = 1.00$) showed significantly greater satisfaction than participants in all the other conditions, $t(92) = 3.05, p = .003, d = 0.64$.⁴ The remaining three conditions did not reliably differ from each other with regard to satisfaction. More precisely, the punishment and feedback about no change condition ($M = 3.76, SD = 1.23$) did not reliably differ from the punishment without feedback ($M = 3.54, SD = 1.00$) or the no punishment no feedback condition ($M = 3.65, SD = 1.21$), $t(92) < 1, p = .546$, and the punishment without feedback condition did not reliably differ from the no punishment no feedback condition, $t(92) < 1, p = .722$.

Intercorrelations of the dependent variables Participants' general affect at time 2 and their satisfaction regarding the interaction were strongly correlated, $r(94) = .493, p < .001$.

Discussion

As predicted, findings from Study 1a revealed that punishers were only satisfied after punishment if it changed the transgressor's attitude. Punishing in itself was neither satisfying nor dissatisfying. Instead, the affective consequences of punishment were dependent on the effect that punishment had on the transgressor, suggesting that people do not have purely retributive motives when they decide to punish.

Moreover, these results demonstrate that it is not the mere presence of feedback that makes punishment satisfying but its content, in particular the feedback entails

⁴ Whenever more than two means are compared to each other, η^2 was calculated by using the η^2 and df values of the contrast ($df = 2$ to 3).

information about change or not. If punishers only wanted transgressors to know why they are being punished (as previous studies have suggested, see Gollwitzer et al., 2011), punishers in Study 1 should have been equally satisfied with both kinds of feedback from the transgressor as both acknowledged the punisher's intent to punish. Yet, the pattern of results from Study 1 suggests that punishers were seeking more than just feedback following the logic of transformative justice motives: they wanted to see a change in the transgressor's attitude.

Participants' general affect decreased in each condition after being treated unfairly, regardless of the experimental manipulation of feedback. Although it correlated with satisfaction related to the interaction partner, general affect did not show the same trend, suggesting that the effect found for satisfaction is not reducible to mood.

Replication Study 1b: Transgressor change affects justice-related satisfaction

The idea of transformative justice motives refers to a deep and true sense of justice restoration through transgressor transformation. In order to make sure that the findings of Study 1a referred to this kind of justice-related satisfaction with punishment and not just general satisfaction about the interaction (e.g., because the offender showed agreement with the participant's decision to punish), because the feedback in the punishment and feedback about no change condition could have been perceived as a challenge to the legitimacy of punishment, the replication study introduced a different satisfaction scale. Study 1b was a slightly modified online version of Study 1a that only focused on the two different feedback conditions (change versus no change).

Method

Justice-related satisfaction scale Different from Study 1, the dependent variable was a 12-item scale (as used in Funk et al., 2014, Study 1) that included items on justice-related satisfaction, deservingness, rumination, and psychological closure items ($\alpha = .85$, order randomized, 1 = not at all, 7 = very much), consisting of the following items: I think that I can now close this chapter, I am content with the way things worked out eventually, I feel satisfied now, I am now able to turn my mind to something else, I think that everybody got what they deserved, I feel relieved, I feel respected, I feel angry, I wish that things would have turned out differently, I feel disappointed, I wonder if I should have acted differently, I think that someone's still got to pay.

Procedure. One hundred and twenty-one participants from the U.S. (38 [31%] female, $M_{age} = 29.3$, $SD_{age} = 10.0$) were recruited on Amazon Mechanical Turk, for its use in social science, (Schmester, Kwang, & Gosling, 2011; Paolacci, Chandler, & Ipeirotis, 2010) to participate in a 10-minute survey on problem solving in virtual groups for a total of \$0.60. Because participants were recruited online, the initial message from the ostensible partner did not refer to fellow students; it read: "You're good at this. Just saw the better we BOTH do, the more bonus we get. Let's go for it!" (typos deliberately added for reasons of credibility). For the sake of time, participants only had one minute to solve up to 15 anagrams, each worth \$0.10 for the team. In addition to the payment that they knew about beforehand, participants also received what they had earned in the game as a bonus payment after the completion of the study.

Results

Preliminary analyses. Eighty participants (66%) decided to reduce their partner's bonus. The data of participants who did not punish were excluded from the analyses ($n = 41$). Those excluded participants who decided not to punish did not differ from participants who decided to punish in terms of gender, $\chi^2(1, N = 121) = 0.34, p = .429$, suspiciousness, $\chi^2(1, N = 121) = 0.27, p = .662$, or the number of solved anagrams, $t(119) < 1, p = .631$. The only significant difference found was that participants who decided not to punish were slightly older ($M = 31.8, SD = 11.0$) than participants who decided to punish ($M = 28.0, SD = 9.3$), $t(119) = 2.00, p = .048, d = 0.38$.

When remaining participants were suspicious that the ostensible partner was not real or had heard about it from other participants, they were equally distributed across conditions, $\chi^2(1, N = 80) = 0.07, p = 1$, they were also excluded from the analyses ($n = 21$). As in Study 1a, the pattern of significant findings would be the same if these suspicious participants were included in the analyses.

The final sample consisted of 59 participants [29%] female, equally distributed across conditions ($M_{age} = 27.4, SD_{age} = 8.6$) with 29 and 30 participants in each condition respectively. Participants solved on average 5.49 anagrams ($SD = 1.87$) and proposed on average 52.90% ($SD = 6\%$, $min = 40\%$, $max = 70\%$) of the bonus for themselves. As punishment, people deducted on average \$0.54 ($SD = \$0.29$) from their partner's bonus. The number of anagram solved and the magnitude of punishment did not correlate with justice-related satisfaction after punishment. The number of anagrams solved, participants' initial allocations, recommendations, and the magnitude of punishment also did not reliably differ between the experimental conditions ($p > .376$).

Manipulation checks. Participants in the change condition agreed more strongly than participants in the no change condition that adjusting their partner's bonus made the partner think about his/her behavior (change $M = 6.69, SD = 0.54$; no change $M = 4.10, SD = 2.30$; Welch's $t(32.3) = 6.01, p < .001, d = 1.54$) and agreed more strongly that adjusting the bonus had caused a change in partner's attitude (change $M = 6.41, SD = 0.78$; no change $M = 2.47, SD = 2.05$; Welch's $t(37.5) = 9.85, p < .001, d = 2.52$).

Effects of transgressor change on general affect and justice-related satisfaction

General affect. Positive affect generally decreased at time 2 ($r = .94$) compared to time 1 ($r = .88$). Different from Study 1, this effect was stronger when the transgressor's feedback indicated change (time 1: $M = 5.98, SD = 0.78$; time 2: $M = 3.32, SD = 1.55$) compared to when the transgressor's feedback indicated no change (time 1: $M = 5.97, SD = 0.84$; time 2: $M = 4.52, SD = 1.50$); $PANAF(1, 57) = 92.74, p < .001, \eta^2 = .619$ condition $F(1, 57) = 6.29, p = .015, \eta^2 = .099$ PANA*condition $F(1, 57) = 8.13, p = .006, \eta^2 = .125$.

Justice-related satisfaction. The main analyses revealed that Study 1 replicated the effect of feedback type found in Study 1, as predicted. Participants who received feedback about change from their ostensible partner expressed significantly more justice-related satisfaction after punishment ($M = 4.23, SD = 1.10$) than participants who received feedback about no change from their partner ($M = 3.51, SD = 1.19$), $t(57) = 2.40, p = .020, d = 0.63$.

Intercorrelations of the dependent variables. Participants' general affect at time 2 and their justice-related satisfaction were strongly correlated ($r(57) = .847, p < .001$).

Discussion

Studies 1a and 1b examined the role of transformative justice motives by manipulating whether transgressors reacted to punishment with a change in attitude or not. Punishment was only satisfying if transgressors changed their attitudes as a consequence of punishment. Because previous studies have demonstrated that people do not punish in order to feel better (Gollwitzer & Bushman, 2012), these findings strongly suggest that punishers hoped to achieve this transformation in the transgressor when they decided to punish.

These findings further suggest that it is less about the presence or absence of general transgressor feedback after punishment, like previous research has suggested, than about what the feedback supposedly signals. Studies that varied whether a transgressor understood that punishment was administered as a consequence for previous behavior (Funk et al., 2014, Study 1; Gollwitzer et al., 2011) might have found effects on satisfaction because participants interpreted such acknowledgments as an indicator of change in the transgressor's attitude. Indeed, making offenders realize why they are being punished is a first step to effecting change in attitude (see also Durkheim, 1893/1964; Miller, 2001; Vidmar, 2001).

Study 1b introduced a new satisfaction scale to make sure that the effects found for transgressor change referred to justice-related satisfaction and do not simply reflect that punishers are more satisfied after transgressor change because transgressors agreed with their decision to punish. The use of a different scale is not a strict experimental test of the possibility that transgressor agreement affected punishers' satisfaction, and future studies might want to manipulate transgressor agreement and transgressor change

separately. Nevertheless, Study 1b could replicate the pattern of results of Study 1a with this newsatisfaction scale, suggesting that even transgressor agreement is an additional source of satisfaction. People felt more justice-related satisfaction, and thus agreed more strongly that everybody got what they deserved, when transgressor change was present compared to when it was not.

There are several limitations to the paradigm used in Studies 1a and 1b that led to the use of a different experimental paradigm in Study 2. First, the experience of unfairness and its reaction to it relied on the use of monetary rewards and monetary punishment. Therefore, the findings might not be ecologically valid for interpersonal transgressions in the real world where the currency of unfair treatment is often not the same as the possible currency of punishment. Second, the operationalization of punishment in Studies 1a and 1b allowed punishers to completely even out the committed injustice. That is, participants could deduct as much money as necessary to make sure that they and their partner would receive about the same bonus in the end. In the real world, punishment behavior might not be able to eliminate the injustice done. Therefore, Study 2 used a different paradigm to further study people's transformative justice motives, one that did not include monetary reward or punishment.

Study 2: Transformative justice motives are distinguishable from deterrent and retributive motives

Results from Studies 1a and 1b do not unequivocally speak to the question of whether punishment necessarily has to deeply transform the transgressor's attitude, the sense of positive prevention (i.e., transformation), or if a change on the behavioral level already suffices (in the sense of negative prevention, i.e., specific deterrence).

because the operationalization of change in Studies 1a and 1b found a change in the transgressor's attitude with the transgressor's change in behavior. Although previous research on punishment severity suggest that people do not punish for deterrent reasons (e.g., Carlsmith, 2006; Carlsmith et al., 2002), the pattern of affective consequences after punishment found in Studies 1a and 1b equally supports deterrent and transformative punishment motives. In order to differentiate transformative justice motives from deterrent punishment motives, Study 2 included a mere behavior change condition in which transgressors did not show a change in attitude but would only indicate a change in their behavior. It was hypothesized that transformative justice motives do not only aim at superficial behavior change but at a profound transformation of the transgressor's attitude that ultimately results in behavior change.

In addition, Study 2 aimed at examining the extent to which inferred suffering contributes to punishers' justice-related satisfaction when they receive feedback about change in the transgressor. It is possible that punishers interpret feedback about change as an indicator for suffering caused by the imposition of punishment. Inferred suffering is not problematic per se, because transgressors who understand that what they did was wrong might experience unpleasant feelings of guilt and therefore suffer (Baumeister et al., 1994). Similarly, it could be that participants think that it is through the experience of suffering that transgressors realize that change is needed and mend their ways (for the idea of punishment as atonement, see Garvey, 1999; for the idea of punishment as secular penance, see Huff, 2001; for the idea of redemption, see McAdams, 2013). Interpreted this way, inferred suffering would be in line with transformative motives.

Inferred suffering could also be in line with retributive motives, however, if punishers interpret it as a restoration of balance. In a retributive sense, suffering that leads to satisfaction could indicate that punishers think transgressors have now received what they deserve (Carlsmith & Darley, 2008; Feather, 1999). Thus, if inferred suffering could explain the effect of transgressor change on participants' justice-related satisfaction findings obtained in Studies 1a and 1b might have been caused by retributive motives and not necessarily by transformative motives. In order to detect a potential connection of transgressor change and inferred suffering and to tease apart their effects on justice-related satisfaction, Study 2 included two versions of the change and no change conditions respectively one that explicitly included suffering and one that did not. In addition, participants rated how much the transgressor presumably suffers as a consequence of punishment.

Furthermore, individual difference measures were added to the experimental procedure to explore how scores on these measures are related to transformative justice motives. First, adding individual difference measures might help to discriminate punishers from nonpunishers. Second, it is possible that the effects of transgressor change on justice-related satisfaction increase or decrease for different personality types. Study 2 exploratorily assessed the following individual difference measures to gain a more distinctive understanding about the generality of transformative justice motives: victim sensitivity, sense of control, personal need for structure, and narcissism.

Victim sensitivity is a subcomponent of justice sensitivity (Schmitt, Neumann, & Montada, 1995). It can be used to assess how people differ in their sensitivity to mean intentions (Gollwitzer & Rothmund, 2009; Gollwitzer, Rothmund, Alt, & Jekel, 2012).

Gollwitzer, Rothmund, & Süßenbach, 2013). First, dispositional differences in victim sensitivity might affect people's desire to effect a change in the transgressor and could explain why some people punish or not. Second, victim sensitive individuals might be more reluctant to be satisfied after transgressors change because they do not trust such a change (Gerlach, Allemand, Agroskin, & Denissen, 2012) or because victim sensitive individuals may not even detect such a change as they focus more on cues related to untrustworthiness than on cues related to trustworthiness (Gollwitzer et al., 2012).

Participants' sense of control (Rotter, 1966; Bandura & Mullen, 1993) was included, first, to examine if it can identify people who decide not to punish, because individuals with a low sense of control could believe that their behavior will not make a difference anyway and refrain from punishing. Second, individuals with a high sense of control might be more satisfied after feedback about transgressor change than individuals with a low sense of control, because their self-efficacy has been confirmed once again.

Personal need for structure (Neuberg & Newsom, 1993) was included because, first, individuals with a high need for structure might be more likely to punish transgressors as they care more about restoring a structure that has been violated. Second, individuals with a high need for structure might be affected more by feedback about transgressor change because it is more important for such individuals that everything is in order. On the other hand, personal need for structure correlates negatively with dispositional forgiveness (Satton, Ward Struthers, & Santelli, 2006). Thus, it is also possible that individuals with a high need for structure are less affected

by feedback about change because it cannot adequately restore the damage that previous transgression has caused to person's perceived structure.

Narcissism was included as measure because it relates negatively with dispositional forgiveness and positively with revenge motivation (Forsyth et al., 2006). It is linked to a retributive and not a restorative orientation toward punishment (Okimoto, Wenzel, & Feather, 2012). Narcissist individuals more easily feel threatened in their self image which increases aggression (Bushman & Baumeister, 1998). Therefore, first, it is possible that individuals with high scores on narcissism are more likely to punish transgressors to boost and restore their self image. Second, individuals with high scores on narcissism might react less pronouncedly to feedback about change than individuals with low narcissism scores because they care more about themselves than about the transgressor, alternately they might react more pronouncedly because they need the feedback about change to boost their self image.

In order to avoid high attrition rates like in Studies 1a and 1b, Study 2 instructed participants to imagine punishing their partner. As discussed earlier, it has been shown that people are very accurate in predicting their own levels of satisfaction after punishment if the presence or absence of transgressor feedback is made salient (Fenk et al., 2014, Study 1).

Moreover, in order to address shortcomings of Studies 1a and 1b, unfair treatment and punishment were in different currencies. Instead of making participants play a game for money, the currency of unfair treatment toward the participant in Study 2 was the time spent working on tasks, and the currency of punishment of the transgressor was the assignment of unpleasant tasks (for the use of unpleasant tasks see,

e.g., Bushman & Baumeister, 1998; Bushman et al., 2001; Gollwitzer et al., 2011, Study 1).

HYPOTHESIS 2:

(A) Transformative justice motives are different from deterrent motives: Punishers are more satisfied after punishment if it results in genuine change that includes a change of attitude and behavior compared to mere behavior change.
(B) Transformative justice motives are different from retributive motives: The effects of change are not reducible to an inferred increase in suffering.

Method

Study 2 was an imagination study conducted in the lab with five experimental conditions manipulated between participants: feedback about change (with or without suffering), feedback about mere behavior change, and feedback about no change (with or without suffering).

Procedure. Participants from the University's psychology subject pool and paid subject pool (N = 205) were recruited to participate in a laboratory study on social behavior in interpersonal situations in exchange of half an hour of course credit or \$8. The study was administered by the computer and asked participants to imagine a hypothetical experiment situation in which they would work together with another person. To make the imagined setup as vivid and as easy to imagine as possible, the instructions consisted of cartoons that displayed the imagined experimenter and the imagined other participant along with screenshots of the imagined experiment instructions.

Participants imagined to participate in an experiment on Office Behavior together with another, unmotivated participant called Tom L. who was introduced saying, 'I want to get back to my TV. I hope this won't take the full 30 minutes.' As explained

by the cartoon experimenter, the participant and the other person would work together on several tasks related to office behavior, and it would be important that participants encounter a variety of different tasks that vary in length and kind. The exact tasks on which the participant and the other person would end up working would be the result of what the participant and the other person decided to choose. As the experimenter explained, there were two steps to this allocation process: First, the two participants would see a grid of six undefined tasks taking between one and ten minutes (e.g., 1, 2, 4, 6, 8, 10 minutes). It would be up to the other person (Tom L.) to decide how long the tasks would take that each person would work on. In a second step, they would see a grid of different task definitions, and it would now be up to the participant to decide what kind of tasks each of them would work during the previously allotted.

Unfair treatment was operationalized by Tom L.'s uneven way of allocating the different tasks that varied in length. Tom L. allocated the tasks in a way that he would only work on three tasks taking seven minutes in total, whereas the participant would need to work on three tasks taking 24 minutes. At this point, participants were asked to report how they would feel right now using similar PANA items to the ones used in Studies 1a and 1b: irritated (–), satisfied, upset (–), happy, angry (–), pleased, annoyed (–), cheerful (.. = .83, order randomized, 1 = not at all, 7 = very much).

Punishment behavior was operationalized through the type of tasks participants decided to allocate to their partner. For that purpose, participants saw a grid of ten different tasks, six of which were designed to be fun (e.g., Watch a brief clip of The Office and rate it, Fun task about the popular TV show, or, Rate the wittiness of comics on Office Behavior, Entertaining), and four of which were designed to

appearing (e.g., Check copy room supply list for types, Somewhat tedious but important, or, Count the number of letters sent last week). Sounds boring? Well, somebody has to take care of it. Participants were asked to assign three tasks to each of them. Because there were six fun tasks total, participants could decide to allocate three fun tasks to themselves and three fun tasks to the other person if they assigned at least one of the four boring tasks to the other person; their allocation was interpreted as punishment behavior (for allocation frequencies, see Results section).

After participants had allocated the tasks, the main manipulation of the experiment was how the other person (Tom) presumably reacted to the participants' task allocations. Participants were randomly assigned to five different feedback conditions that varied in the cartoon wording of Tom's reaction. Participants in the no change conditions [with suffering] read, 'Hey, you allocated almost all of the boring tasks to me. [That will really suck!] Okay, I was kind of unfair with the timing but I don't see what's wrong with that. I really want to go back to my TV show.' Participants in the mere behavior change condition read the same that participants in the no change suffering condition read followed by 'But if that makes you treat me bad, I wouldn't do it again next time. Get it....' Participants in the change conditions [with suffering] read, 'Hey, you allocated almost all of the boring tasks to me. [That will really suck!] Okay, I kind of was unfair with the timing... and now I see what's wrong with that. I wouldn't do it again next time. I understand.. See Appendix for the screen shots of the feedback

Next, the main dependent variable was the justice-related satisfaction scales used in Study 1b (12 items... = .83, order randomized).

As a first manipulation check of perceived change, participants indicated to what extent they thought their allocation has caused a change in Tom L.'s attitude that Tom L. regrets his previous allocation of task duration that Tom L. thinks that what he did was wrong, and to what extent Tom would probably act the same way again in the future (recoded), ... = .66 (order randomized, 1 = not at all, 7 = very much). As a second manipulation check for perceived suffering, participants were also asked to what extent they thought the other participant was suffering as a result of how they allocated the tasks (1 = not at all, 7 = very much).

Next, as potential covariates, participants indicated how important it would be to them that the other participant changes his attitude, that the other participant changes his behavior in the future, and that the other participant suffers as a result of how allocated the tasks (order randomized, 1 = not important at all, 7 = very important).

Participants then started working on one task they watched a short video clip from the TV show, *The Office*, and answered some filler tasks. Next, they indicated how they would feel during the remainder of the study, working on their tasks for 24 minutes, using the same PANA items that were assessed earlier (... = .92). At the end of the imagination part of the study, participants were asked to imagine that the other participant was done after seven minutes but had to wait for the experimenter who came back when they were done after 24 minutes.

Next, participants filled out various individual difference measures. First, participants filled out the 10-item Victim Sensitivity Scale (Schmitt, Gollwitzer, Maes, & Arbach, 2005, e.g., I ruminate for a long time when other people are being treated better than me, ... = .87). Second, participants' sense of control was assessed using four

items from Lachman & Weaver, 1998, e.g., Whether or not I am able to get what I want is in my own hands $r = .71$). Third, personal need for structure was measured (12 items from Neuberg & Newsom, 1993, e.g., I hate to be with people who are unpredictable $r = .80$). Fourth, participants filled out the 18-item Narcissistic Admiration and Rivalry Questionnaire (NARQ, Back et al., 2013, e.g., I enjoy it when another person is inferior to me $r = .78$). See Appendix for a list of all items used.

At the end of the study, participants answered some demographic items, were thanked and debriefed about the purpose of the study, and gave their fully informed consent.

Results

Preliminary analyses. 63 participants (31%) assigned no boring tasks to their partner and were excluded from the final analyses⁵, resulting in a final sample $N = 142$ (91 [64%] female) with 2729 participants in each condition, out of which 40 (28%) assigned one boring task to Tom L., 34 (24%) assigned two boring tasks, and 68 (48%) assigned three boring tasks to Tom, equally distributed across the five conditions, $\chi^2(4, N = 142) = 350, p = .900$. The number of boring tasks assigned did not correlate with the dependent variable $r(140) = .095, p = .261$. Similarly, if treated as a categorical variable, the number of boring task assigned did not affect the dependent variable, $F(2, 139) = 1.93, p = .149$. The main analyses therefore collapse across the different numbers of boring tasks assigned to the partner: the magnitude of punishment

There were disproportionally more female participants in the behavior change condition and the two change conditions than in the two no change conditions,

⁵ Including these participants in the final analyses would not alter the pattern of findings reported below.

$F(4, N = 142) = 14.4, p = .006$ However, participant gender had no effect on justice related satisfaction, $F(1, 141) = 0.09, p = .762$. The results reported below were therefore not controlled for participant gender.

None of the items that measured participants' self-assessment of the importance of change in the transgressor's attitude or behavior and of the importance of transgressor suffering (assessed as covariates) differed between experimental conditions, $F_s < 1.51$, $p_s \geq .204$. These items also did not correlate with the dependent variables, $r_s \leq .075$, $p_s \geq .373$. They were therefore not controlled for in the analyses reported below.

Manipulation checks.

Perceived change in the transgressor's assessment of transgressor change significantly differed between the experimental conditions, $F(4, 137) = 4.22, p = .003, \eta^2 = .110$. Planned contrast analyses revealed that the two change conditions (with and without suffering) differed significantly from the two no change conditions (with and without suffering), $t(137) = 4.02, p < .001, d = 0.69$. (This would also hold true if one only compared the no change no suffering condition, $M = 3.85, SD = 1.21$, with the change no suffering condition, $M = 4.53, SD = 1.01$, $t(137) = 2.31, p = .022, d = 0.62$ and the no change suffering condition, $M = 3.70, SD = 1.19$, with the change suffering condition, $M = 4.69, SD = 1.14$, $t(137) = 3.38, p = .001, d = 0.85$) Ratings for the mere behavior change condition lay in between the no change and the change conditions ($M = 4.12, SD = 0.97$).

Importantly, the single item asking if Tom L. would probably act the same way again in the future did not significantly differ between the behavior change condition ($M = 4.66, SD = 1.45$) and the change no suffering condition ($M = 4.55, SD =$

1.55), $p = .811$, indicating that feedback in the mere behavior change condition was indeed perceived as an indication of intended behavior change. The mere behavior change condition differed significantly from the change suffering condition [$M = 3.71$, $SD = 1.49$], $p = .032$, but that condition was excluded from the final analyses for other reasons (see next paragraph.)

Perceived suffering Analyses of variance (ANOVA) revealed that differences between participants' perceptions of the transgressor's suffering were marginally significant, $F(4, 137) = 2.39$, $p = .054$, $\eta^2 = .065$. Yet, a closer look at the means indicated that the manipulation of transgressor's suffering only led to higher suffering ratings for the no change conditions (no suffering $M = 3.00$, $SD = 1.71$; suffering $M = 4.28$, $SD = 2.12$; $t(137) = 2.60$, $p = .010$, $d = 0.66$) but not for the change conditions (no suffering $M = 4.00$, $SD = 1.96$; suffering: $M = 3.25$, $SD = 1.69$, $t(137) = -1.55$, $p = .125$, $d = -0.41$). Suffering ratings for the mere behavior change condition lay in between ($M = 3.41$, $SD = 1.62$). Importantly, participants' ratings of perceived suffering did not correlate with justice-related satisfaction, $r(140) = -.005$, $p = .955$. Because the manipulation of perceived suffering did not work, the main analyses for the dependent variables excluded the two suffering conditions, $N = 85$.

The effects of transgressor change on general affect and justice-related satisfaction.

General affect. There was a general finding that participants felt better at time 2 ($M = 4.24, SD = 1.34$) than at time 1 ($M = 2.17, SD = 0.76$), $F(1, 82) = 164.19, p < .001, \eta^2 = .667$. There were no differences between experimental conditions ($F(2, 82) = 1.69, p = .191$), and no differential effects for the different experimental conditions, $PANA \times \text{condition } F(2, 82) = 1.60, p = .209$.

Justice-related satisfaction. Although ANOVA results did not yield significance, $F(2, 84) = 2.23, p = .114, \eta^2 = .052$, planned contrast analyses revealed as predicted that participants in the change condition ($M = 4.89, SD = 1.12$) were significantly more satisfied than participants in the no change condition ($M = 4.34, SD = 0.87$), $t(82) = 2.05, p = .044, d = 0.55$, and values for participants in the mere behavior change condition lay in between ($M = 4.51, SD = 1.00$).

The same pattern of findings would emerge if the ANOVA included both suffering conditions ($N = 142$) and if contrasts collapsed across the two suffering and no suffering versions of the change and no change conditions respectively (within the change and no change conditions the suffering and no suffering versions do not significantly differ from each other).

Intercorrelations of the dependent variables. Participants' general affect at time 2 and their justice-related satisfaction were moderately correlated ($r(83) = .281, p = .009$).

Effects of individual difference measures. None of the individual difference measures differed between the experimental conditions ($F(4, 137) < 1.01, ps \geq .403$), and none of them correlated with justice-related satisfaction ($r(140) < .134, ps \geq .112$).

Of all of the scales (including their subscales), only narcissism ($M = 2.90, SD = 0.59$) and victim sensitivity ($M = 4.77, SD = 1.06$) correlated with participants' punishment magnitude (range of number of boring tasks assigned to the other, $M = 1.52, SD = 1.24$), indicating that people who are more sensitive toward victimization ($r(203) = .138, p = .048$) and people with higher narcissism scores ($r(203) = .161, p = .021$), assign more of the boring tasks to their partners. As for narcissism, the correlation seemed to be driven by the NARQ subscale rivalry which correlated with number of boring tasks assigned to the other ($r(203) = .203, p = .004$), but not by the subscale admiration which was not related to number of boring tasks assigned to the other ($r(203) = .055, p = .431$).

Analyses of covariance (ANCOVA) testing for interactions between experimental condition and the individual difference measures revealed only a significant interaction of experimental condition with personal need for structure (PNS) condition $F(2, 79) = 4.32, p = .017, \eta^2 = .099$; PNS scale $F(1, 79) = 0.60, p = .440$; condition * PNS scale $F(2, 79) = 3.83, p = .026, \eta^2 = .088$. Descriptive statistics using medians split show that the interaction indicates that people who have a lower need for structure are more satisfied after a change compared to no change ($d = 0.76$) than participants who have a higher need for structure ($d = 0.36$), see Figure 1 for the pattern of means.

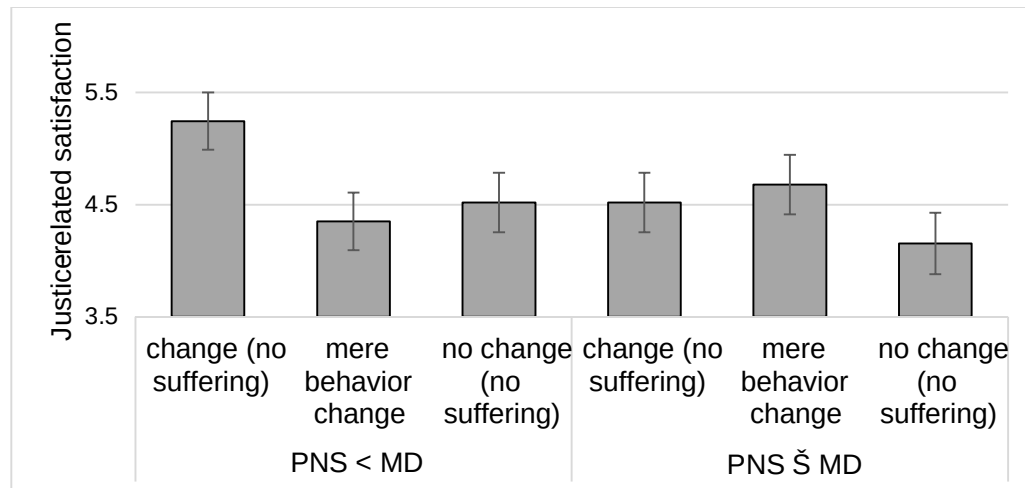


Figure 1. The effect of different conditions of transgressor change on justicerelevant satisfaction in Study 2 split for participants' scores on personal need for structure (PNS) below or above the median

Discussion

Using a different experimental paradigm than before, Study 2 replicated the effects found in Studies 1a and 1b and demonstrated once again that punishers are more satisfied after punishment if the transgressor has changed compared to when the transgressor has not changed. Thus, the effect of transformative justice motives is generalizable beyond monetary game contexts and also occurs when unfair treatment and punishment are in different currencies. As change in the transgressor left punishers satisfied although they could not even out the unfair situation and imagined at that time that they would have to work on tasks that took longer than their partners. These findings suggest that justicerelevant satisfaction is more than just satisfaction about the tangible outcome of an interaction and refers to an inner sense of restored justice.

Results from Study 2 suggest that transformative justice motives go beyond deterrent motives and highlight the importance of paying more attention to the psychological difference between motives related to positive prevention and negative prevention. Punishers were only significantly more satisfied if the transgressor indicated a change in his attitude and behavior, but not if he merely indicated a change in his behavior. Thus, punishers seem to have aimed at a genuine transformation in the transgressor and did not only hope to shape his behavior for the better. These findings suggest that transformative justice motives have prosocial aspects and seem to include reintegrative features that target more than the transgressor's visible behavior.

Although participants could only punish Tom L. by assigning unpleasant tasks, correlational findings suggest that it was not mainly the unpleasantness participants hoped to impose on the transgressor that made punishment satisfying. Perceived suffering did not correlate with justice-related satisfaction, rendering retributive explanations unlikely. Findings from Study 2 therefore strongly suggest that the allocation of unpleasant tasks was not used as an end in itself, but in a retributive way. Instead, the allocation of unpleasant task served as tool to achieve a transformation in the transgressor.

A potential objection to this interpretation is that, for the no suffering conditions that were ultimately used for the main analysis, participants rated the transgressor's suffering in the change condition as higher than in the no change condition. However, as previously mentioned, the perception of suffering did not correlate with justice-related satisfaction overall. Moreover, the change and the no change conditions showed the same pattern of effects on justice-related satisfaction for the suffering and no suffering

versions. The no change condition with suffering even received higher ratings on suffering than the change condition with suffering but actually showed lower ratings on justice-related satisfaction. These results suggest that transformative motives are not reducible to and different from retributive motives.

Future studies should use different ways to manipulate the presence or absence of suffering, as Study 2 had to exclude the two suffering conditions from the main analyses after manipulation checks did not confirm a successful manipulation. Such studies could further examine the difference between retributive and transformative motives by including feedback conditions that do not only manipulate the presence or absence of suffering but also the presence or absence of joy. It remains to be tested if a happy transgressor who changes leads to the same findings as an indifferent transgressor who changes.

General affect in Study 2 showed a surprising pattern indicating an increase in positive mood at time 2. In hindsight, the choice of time 2 was ideal as participants had just watched a funny short clip of a famous comedy TV show (The Office). Therefore, the interpretability of this finding is limited. Future studies need to choose the different time points that assess general affect more carefully.

Individual difference measures were included for exploratory reasons so if they should be interpreted with caution. It is possible that the effects found are statistically speaking false positives and just a result of a cumulative Type I error due to numerous tests. Nevertheless, these findings are a first step to better understand the effect that transgressor change has on punishers' justice-related satisfaction. The correlations found for victim sensitivity and narcissism with punishment magnitude

conceptually replicate previous findings on the link between victim sensitivity and a focus on cues related to trustworthiness (Gollwitzer et al., 2012) and on the link between narcissism and aggressive tendencies (Bushman & Baumeister, 1998).

The interaction finding of transgressor change and participants' personal need for structure on participants' justice-related satisfaction needs to be interpreted with caution for the same reason mentioned before as well as for statistical power issues. The means in Figure 1 illustrating the interactive pattern are based on a small sample size in each cell ($n \sim 14$). Nevertheless, the interaction findings are in line with previous findings on the negative correlation between personal need for structure and dispositional forgiveness (Eaton et al., 2006). Individuals with high need for structure are less prone to forgive transgressors, and therefore they do not react to transgressor change with the same degree of satisfaction that individuals with a lower need for structure indicate.

More research is needed to better understand the exact role of personality variables in the context of transformative justice motives. If the interaction of change and personal need for structure can be replicated in future studies, the results might demonstrate the importance of punishment, *fit*. Some people might punish for transformative motives, whereas others (in this case people with high personal need for structure) might not. Alternately, these results could mean that people with high personal need for structure need more indicators of change to fully trust a transgressor's apparent transformation.

Study 2 used an imagination paradigm to avoid losing too many subjects for the final analyses. Still, the punishment rate was just 69%, which seems that even if studies are

designed to elicit punishment behavior, a substantive number of individuals decide not to punish. If one wants to fully understand transformative justice motives, it seems to be indispensable to study both punishers and nonpunishers. For Studies 3 and 4 different experimental designs were used in order to include participants in the main analyses who decide not to punish. This approach did not only have the advantage of increasing the final sample size but also offered the possibility to examine interesting questions, such as Do punishers and nonpunishers differ in their reaction to transgressor change? What are the different strategies besides punishment that people use to react to unfair treatment and what do they aim at by doing so?

Study 3: Transformative justice motives aim at

authentic change on affective, behavioral, and cognitive levels

The main focus of Study 3 was to examine the importance of perceived authenticity of transgressor change by manipulating its affective component. Genuine change is a composition of affective, behavioral, and cognitive components. As Study 2 demonstrated, change that includes both a behavioral component (assuring not to repeat the transgression) and a cognitive component (understanding that the behavior was wrong) led to more satisfaction than change that consisted solely of a behavioral component (assuring not to repeat the transgression without understanding that the behavior was wrong). Study 3 also added a manipulation of the affective component by manipulating indicators for the transgressor's experience of remorse.

Remorse is a troubling feeling of distress caused by a sense of guilt for past wrongs (see, e.g., Slovenko, 2006). Previous research has found that transgressors who are perceived to experience remorse are rated to be less likely to reoffend (Gold &

Weiner, 2000; Pipes & Alessi, 1999). Remorse communicates that transgressors still hold on to community values and are less likely to repeat a transgression (Soffman, 1971; Lazare, 2004; Proeve & Tudor, 2010).

For Study 3, it was hypothesized that the perception of remorse experienced by the transgressor after punishment serves as an indicator of authentic change. Only if change is perceived to be authentic, that is if the displayed affective component changes in line with the displayed behavioral and cognitive components, it should affect justice-related satisfaction. To study the match or mismatch of indicators of affective components with feedback about the behavioral and cognitive components change, Study 3 added facial stimuli to the written transgressor feedback that were created to look more or less remorseful.

In addition, Study 3 examined the question of whether it is only punishers who have transformative justice motives or if non-punishers have them, too. On the one hand, it is possible that only people who have transformative justice motives punish. In this sense, transformative justice motives would be a sufficient condition for punishment. If transformative justice motives emerge, it would guarantee that people punish. On the other hand, it is possible that transformative justice motives generally arise after witnessing transgressions, and only some people express transformative justice motives through punishment.

In order to examine how general or specific transformative justice motives are, Study 3 was again a study about actual (not imagined) behavior. The design from Study 2 was slightly changed to make the apparent behavior of the partner more realistic and less staged. In order to operationalize unfair treatment, ostensible partners allocated actual

tasks to themselves as well as to the participants instead of allocating different task durations. Unfair treatment by the transgressor and the participant's reaction were therefore again in the same 'currency' which was the factor of the tasks.

After being assigned to work on unpleasant tasks, it was easy for participants to either allocate the remaining tasks fairly in return or to also behave unfairly (by constituting the nonpunishers and punishers respectively). Although this naturalistic approach includes the possibility that participants' punitive reactions are also related to compensatory notions of justice, participants might assign themselves fun tasks because none were allocated to them from their partner, the reaction nevertheless serves as indicator for punitive notions of justice by deciding to allocate all of the fun tasks to themselves, participants equally decide to allocate the remaining unpleasant tasks to their partners.

Lastly, Study 3 started to examine the effects of justice-related satisfaction on real-world consequences such as future behavior toward the transgressor. For that purpose, participants were asked to indicate their willingness to interact with the interaction partner again in the future.

HYPOTHESIS3:

Change in the transgressor needs to be perceived as authentic in order to justice-related satisfaction. Such authenticity refers to a composition of affective (feeling remorseful), behavioral (not repeating the transgression), and cognitive (understanding the wrongfulness) components of change

Creation and validation of facial stimuli

The affective component of change was manipulated by facial information that the transgressor could add to the written feedback. In order to create the facial stimuli the specific facial information that underlies social judgments of remorse as first

identified by having participants rate frontal renderings of 3D face scans from the Basel Face Model (see Paysan, Knothe, Amberg, Romdhani, & Vetter, 2009). Subsequently, following Walker and Vetter (2009) face space method, the derived information about remorseful appearance was transferred to novel 2D photographs to create remorseful and remorseless-looking faces.

Identifying the look of remorse. Participants ($N = 544$, 293 male, 251 female; $M_{\text{age}} = 32.0$, $SD_{\text{age}} = 11.9$) were recruited from the U.S. and were paid \$0.40. They saw a random subset of 15 out of 153 male and female 3D scans (10 for each face ~ 55°). For each face, participants were asked to imagine that this person had committed a crime and is now facing trial. They then indicated to what extent they thought the person feels genuine remorse, feels guilty and regrets the crime, feels truly sorry for the victim, wants to make amends for the harm caused, and knows that the behavior was wrong (for each face: 1 = not at all, 7 = very much; fixed order). The computer program used the mean of these ratings for each face to identify the physical information that participants relied on when they inferred remorse from this information (i.e. the visual essence of remorse) was applied to novel faces to create remorseful and remorseless-looking versions of the same individuals for the method see Walker and Vetter (2009).

Pretesting the computer-generated faces. Two male and two female Caucasian facial identities with neutral expressions were used from the Radboud Faces Database (Langner et al., 2010) to create both remorseful and remorseless versions. Together with the neutral versions, they were pretested in order to identify one male and one female

facial identity whose appearance lead to significantly different perceptions for the remorseful and remorseless versions.

For that purpose, American participants ($N = 480$ [234 male], $M_{age} = 32.0$, $SD_{age} = 10.9$) were recruited on Mturk to participate in a 1-minute survey on their opinion on one face and were paid \$0.10. Thirteen participants failed an attention check item at the beginning of the survey reading, "To show that you are alert, please leave the below item blank and skip to the next page. Are you reading this item carefully?" (1 = not at all, 7 = very much), and were excluded from the analysis (see Oppenheimer, Meyvis, & Davidenko, 2009). Participants were randomly assigned to rate the appearance of one of the twelve facial stimuli (remorseless, neutral, or remorseful for one of the two male and two female facial identities). Participants were asked to which extent they think a person feels genuine remorse, feels guilty and regrets the crime, feels truly sorry for the victim, wants to make amends for the harm caused, and knows that his/her behavior was wrong, and the averaged ratings were analyzed separately for each of the facial identities.

The two faces that were selected for Study 3 (see Figure 2, page 51) significantly differed in their ratings for the remorseful versus remorseless versions of a facial identity. More precisely, ratings for the three versions of the male face (Raf090_05, 118; ... = .93) sufficiently differed on the remorse rating ($F(2,115) = 6.83$, $p = .002$, $\eta^2 = .106$, indicating that the version with the manipulated remorseless expression ($M = 2.87$, $SD = 1.09$) differed significantly from the version with the manipulated remorseful expression ($M = 3.89$, $SD = 1.35$), $p < .001$, and the original neutral picture was rated in between ($M = 3.29$, $SD = 1.24$). Ratings for the three versions of the female face

($R^2 = .95$), also sufficiently differed for the three versions ($F(2, 114) = 6.19, p = .003, \eta^2 = .098$). The original neutral picture ($M = 3.68, SD = 1.14$) was rated very similar to the remorseless expression ($M = 3.60, SD = 1.48$) but the manipulated remorseless expression differed significantly from the version with the manipulated remorseful expression ($M = 4.53, SD = 1.23, p = .002$), as desired.

Method

Study 3 was an online study that experimentally varied facial expression that went along with the feedback transgressors gave to participants to assess its effect on participants' justice-related satisfaction. In total, there were five experimental conditions manipulated between subjects: feedback about change (paired with a remorseful or a neutral face), feedback about behavior change (paired with a neutral face), and feedback about no change (paired with either a neutral or remorseless face).

Procedure. Five hundred Mturk workers from the U.S. (266 [53%] male, $M_{age} = 32.5, SD_{age} = 10.9$) were recruited online to participate in a 10-minute study on office behavior and were compensated with \$1. At the beginning of the study, they were told that they would work on several office-related tasks with another Mturk worker. They entered a nickname for the study and indicated their age and gender to be paired with a Turker who is similar to them. Participants always interacted with our pre-programmed web interface. Male participants were led to believe that the other Mturker they were paired with was Tony, female participants were led to believe that the other Mturker they were paired with was Tina. Participants who did not want to indicate their gender ($n = 2$) also filled out the female version of the experiment

Before the study was supposed to start, participants were asked to send a message to their fellow Mturker and, to familiarize them with the use of pictures they were asked to pick a picture of a face that best described how they feel right now that would go along with the message. The picture always the facial identity of their indicated gender and possible choices were the happy, neutral or contemptuous versions taken from the Radboud Face Database and the manipulated remorseful version. Participants then received the message that their fellow Mturker (Tina or Tony respectively) had supposedly written, reading hi [name of participant]. hope this won't take too long, paired with a neutral facial expression (female or male respectively)

Participants were told that the study examined how people use the computer to work on several tasks related to office behavior and that their decision times for each task would be recorded. Every participant would work on four tasks that would take about three minutes in total. There would be different kinds of tasks and the two of them would decide together who will work on which tasks. First, Tony/Tina would allocate two tasks to each of them. Next, the participant would allocate two tasks to each of them. Once they had decided who will work on which tasks, they would both work on their particular tasks simultaneously.

First, participants indicated how they felt using some of the PANA items used in previous studies (happy, annoyed (-), pleased, upset (-), angry (-), satisfied ... = .83, order randomized, 1 = not at all, 7 = very much).

Next, participants could see how Tony/Tina was asked to allocate tasks to each of them. In the list of the four available tasks, there were identical fun ones (Watch a short clip of •The Office and rate it, Fun and entertaining (versions 1 and 2) and two

identical boring ones (,Highlight office-related words in a text Be sure not to miss one), versions 1 and 2) Tony/Tina always allocated the two boring tasks to the participant and the two fun tasks to himself/herself. As an operationalization check for perceived unfairness, participants were asked to indicate how much they are looking forward to do the tasks that the other Turker has picked for them (1 =not at all, 7 =very much).

Next, participants could allocate two tasks to each other. The list of possible tasks included again two identical fun tasks (,Rate comics on Office Behavior Entertaining and funny versions 1 and 2) and two identical boring tasks (,Check copy room supply list for typos, Somewhat tedious but important versions 1 and 2) allowing participants to punish their interaction partners or to treat them fairly.

Before participants were supposed to work on the tasks they picked, they and the ostensible other Turker could simultaneously send one last message to each other along with a picture of their indicated gender that best described how they felt. This time, possible choices were the neutral and contemptuous versions taken from the Radboud Faces Database, along with the manipulated remorseful and remorseless versions. In order to standardize participants' levels of emotional expression, previous research suggests that the possibility to express negative emotions affects people's punishment behavior (Xiao & Houser, 2005), the webpage was programmed to automatically advance after nine seconds, followed by an error message that the message could not be sent. Participants were then randomly assigned to five different feedback conditions.

Figure 2. From left to right: remorseful version, neutral version (original picture) and remorseless version of the face that went along with the feedback message in Study 3 (female identity Rafd090_26 and male identity Rafd090_05)

The presence or absence as well as the authenticity of transgressor change was operationalized by the wording of Tony's/Tina's message and the facial expression that they apparently picked to go along with that message. The two conditions (paired with either a remorseful face or a neutral face) read, "got it, that wasn't very nice of me, I shouldn't have been so selfish, I would do it differently now if I could." The mere behavior change condition (always paired with a neutral face) read, "got it, you think that wasn't very nice of me, I sometimes gotta be selfish, I guess, I would do it differently."

now if i could, if that's what someone expects from me. but not because, i and the no change conditions (paired with either a neutral face or a remorseless face) read, got it, you think that wasn't very nice of me? sometimes i gotta be selfish? would do the same thing again?,, for faces see Figure 2

As a first manipulation check, right after participants saw their partner's message, they were asked to indicate how they think the other Turker is feeling right now on a list of emotions that included remorseful (1 = not at all, 7 = very much).

Before the apparent main tasks were supposed to start, participants were asked to answer some questions about the interaction with the other Turker, constituting the main dependent variable just related satisfaction as used in previous studies (.12, $r = .79$, order randomized, 1 = not at all, 7 = very much), as well as the item, 'Would you like to interact with the other Turker again in the future?' (1 = not at all, 7 = very much). After these items, there was an attention check (see Oppenheimer et al., 2009) to screen out participants who did not read the items carefully. Are you reading the items carefully? If yes, mark 1, if not, mark 7.

Participants then read that the interactive part of the experiment has ended now and that, for the sake of time, they will only need to work on one task related to office behavior instead of four. All of the participants subsequently rated three office comics in terms of how funny and realistic they are. Next, participants indicated again how they felt, using the same PANAS items that were used at the beginning of the experiment ($r = .84$).

At the end of the study's second manipulation check, participants answered some questions about the task allocations by themselves and the other Turker.

extent they thought their task allocation has caused a change in the other Turkers' attitude, the other Turker regrets her/his allocation of tasks, and the other Turker would probably act the same way again in the future (recoded). (1 = not at all, 7 = very much). Participants were asked to briefly describe why they allocated the tasks the way they did in a text box.

In order to get as many data as possible about participants' consciousness about the experimental paradigm, participants were also asked to indicate on three continuous items what their thoughts were about whether the other Turker was real or not (1 = definitely real, 5 = definitely fake), how often they doubted if the other Turker was real or not during the study (1 = never, 5 = all of the time), and how much these doubts affected their behavior toward the other Turker (1 = not at all, 7 = very much).

Participants were asked to provide some demographic information and described how they found the study, for instance if they were searching for a certain term on Mturk, or if they had read about the study on forums. They were thanked, debriefed about the deception used, and asked for their fully informed consent at the end of the study. Lastly, they indicated if they had heard about the true purpose from other Mturkers before they participated, and whether they had known from other people that the partner was not real (both yes/no).

Results

Preliminary analyses. Participants who allocated two boring tasks to themselves ($n = 74$), who did not pass the attention check ($n = 12$), who indicated that they had known from other Mturkers that the partner was not real or who had heard about the true purpose of the study from other Mturkers ($n = 16$), or who did not give informed consent

and wanted their data removed from the dataset (3) were excluded from the analyses, leaving a final sample of 497 participants (n in each condition = 74-84; 184 female [46%], $M_{\text{age}} = 32.0$, $SD_{\text{age}} = 10.7$), out of which 297 (75%) identified as white/European American. Participants' self-reported ethnicities did not affect the findings. Moreover, analyses that look at the complete dataset ($N = 497$) would lead to the same patterns of findings reported below.

Male participants (57%) were significantly more likely than female participants (47%) to allocate two boring tasks (compared to just one boring task) to their partner ($t(397) = 6.24$, $p = .044$). But gender was equally distributed across conditions ($t(397) = 11.31$, $p = .185$), did not significantly correlate with justice-related satisfaction ($r(395) = .092$, $p = .067$), and including gender or task allocation in the main analyses did not alter the results reported below.

Means of the continuous suspicion items (e.g., asking if participants doubted if their partner was real or not) did not differ between experimental conditions ($F(1, 392) < 1.33$, $ps \geq .260$). In addition, participants' thoughts about whether the other Turker was real or not did not correlate with the dependent variable ($r(395) = .045$, $p = .375$), neither did participants' assessments of how often they doubted if the other Turker was real or not, ($r(395) = -.067$, $p = .183$). Participants' ratings about how much their doubts affected their behavior toward the other Turker significantly correlated with justice-related satisfaction ($r(395) = -.167$, $p < .001$). Including this item as covariate did not lead to different findings. Similarly, including participants' ratings of how much they are looking forward to do the tasks that the other Turker has picked for them ($r(396, SD = 1.80)$) did not alter the significant findings reported below.

Manipulation checks. Participants' ratings on how remorseful the other partner felt differed significantly between experimental conditions, $W(4, 195.5) = 45.41$, $p < .001$, $\eta^2 = .306$. Planned contrast analyses revealed that participants in the two change conditions (remorseful face $M = 4.40$, $SD = 1.87$; neutral face $M = 4.39$, $SD = 1.76$) differed significantly from participants in the two no change conditions (neutral face $M = 2.00$, $SD = 1.34$; remorseless face $M = 1.97$, $SD = 1.28$) $W(1) = 13.69$, $p < .001$, $d = 2.10$, and that both the no change and the change conditions differed significantly from the behavior change condition (neutral face $M = 3.13$, $SD = 1.79$) $W(1) = 4.99$, $p < .001$, $d = 0.91$, and $W(1) = 5.41$, $p < .001$, $d = 0.83$, respectively. The kind of face that was paired with the feedback did not make a difference as to how remorseful a partner was perceived in the change conditions, remorseful versus neutral face $W(1) = 164.1$, $p = .987$. Similarly, there were no significant differences in remorseful ratings between neutral and remorseless faces of the no change conditions, $W(1) = 150.0$, $p = .899$.

Participants' averaged ratings about whether their task allocation has caused a change in their partner's attitude, that their partner regrets the allocation of tasks, and that their partner would probably not act the same way again in the future differed significantly between experimental conditions, $F(4, 392) = 107.88$, $p < .001$, $\eta^2 = .524$. Planned contrast analyses revealed that participants in the two change conditions (remorseful face $M = 4.86$, $SD = 1.28$; neutral face $M = 4.81$, $SD = 1.23$) differed significantly from participants in the two no change conditions (neutral face $M = 2.07$, $SD = 1.11$; remorseless face $M = 2.05$, $SD = 1.02$) $t(392) = 20.76$, $p < .001$, $d = 2.10$, and that both the no change and the change conditions differed significantly from the

mere behavior change condition (neutral face $M = 3.40$, $SD = 1.29$) $t(392) = 8.05$, $p < .001$, $d = 0.81$, and $t(392) = 8.80$, $p < .001$, $d = 0.89$, respectively. Again, the kind of face that was paired with the feedback did not make a difference within the change or the two no change conditions as to how much participants agreed to the three items on perceived change, change conditions $t(392) < 1$, $p = .787$, no change conditions $t(392) < 1$, $p = .887$.

Effects of transgressor change on general affect, justice-related satisfaction, and willingness to interact again.

General Affect. The change in general positive and negative affect measured at the beginning and at the end of the study was moderated by experimental condition, $PANA \times \text{condition}$ $F(4, 392) = 2.64$, $p = .034$, $\eta^2 = .026$, indicating that general affect increased for the change conditions but decreased for the no change conditions, see Figure 3 for changes in means from time 1 to time 2. Main effects for repeated measurement (within participants) or experimental condition (between participants) were both not significant; $PANA(1, 392) = 1.70$, $p = .193$; condition $F(4, 392) = 0.93$, $p = .445$.

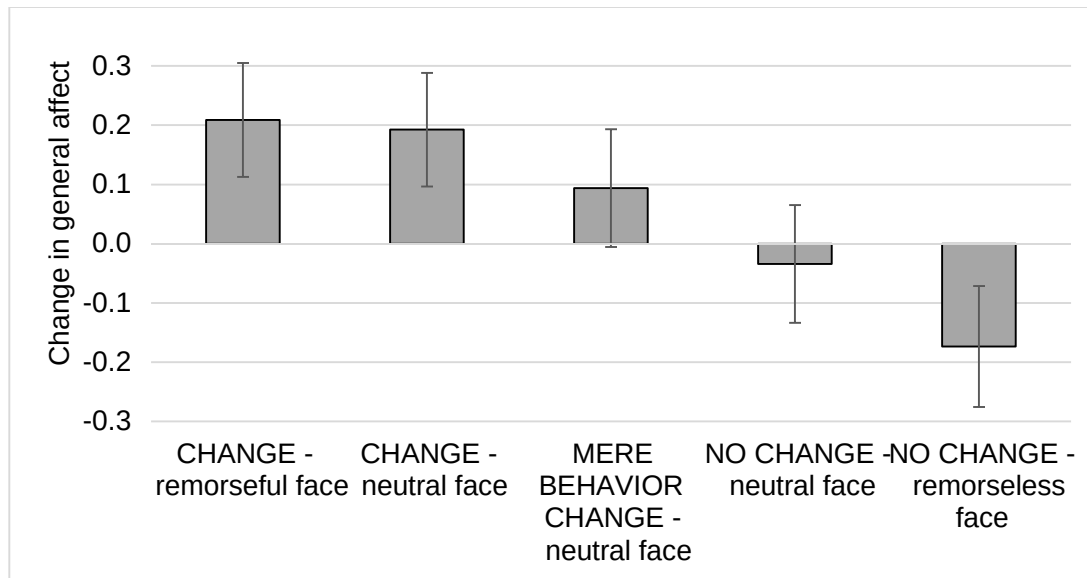


Figure 3. Change in general affect for different experimental conditions in Study 3 (positive values indicate more positive general affect at time 2 than at time 1)

Justice-related satisfaction As predicted, participants' justice-related satisfaction scores differed significantly between the different experimental conditions $F(4, 392) = 2.86, p = .023, \eta^2 = .028$. Planned contrast analyses replicated the general effect that the two change conditions led to greater justice-related satisfaction than the two no change conditions, $t(392) = 3.06, p = .002, d = 0.31$. A closer look revealed that only the authentic change (i.e. change paired with a remorseful face) and authentic no change conditions (i.e. no change paired with a remorseless face) differed significantly from each other, $t(392) = 3.15, p = .002, d = 0.49$. The change and no change conditions that were presented with a neutral face did not significantly differ from each other, $t(392) = 1.16, p = .248, d = 0.19$. For means see Figure 4. Like in Study 2, justice-related satisfaction for the mere behavior change condition lay in between the change and the

no change condition, replicating that effects of indicators for a transformative change are different from effects of indicators for mere behavior change.

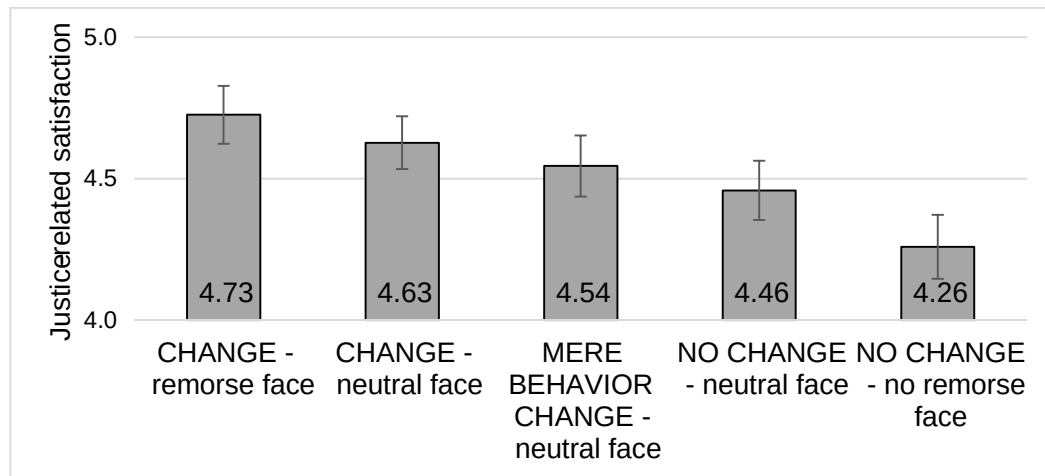


Figure 4. Means of justice-related satisfaction for each of the experimental conditions in Study 3

Willingness to interact again Transgressors feedback about the presence or absence of change significantly affected participants' willingness to interact with them again, $F(4, 392) = 3.76, p = .005, \eta^2 = .037$. Planned contrasts revealed that participants in the two change conditions (remorseful face $M = 3.88, SD = 1.86$; neutral face $M = 3.88, SD = 1.88$) were significantly more willing to interact with the other Mturker again in the future than participants in the two no change conditions (neutral face $M = 3.24, SD = 1.94$; remorseless face $M = 2.92, SD = 2.01$), $t(392) = 3.74, p < .001, d = 0.38$. This difference between change and no change was significant for both the authentic conditions with a matching face, $t(392) = 3.15, p = .002, d = 0.50$, and the inauthentic conditions with neutral faces, $t(392) = 2.13, p = .034, d = 0.33$. Ratings for participants

in the mere behavior change condition lay in between (neutral face $M = 3.41$, $SD = 1.86$).

Intercorrelations of the dependent variables Participants' general affect at time 2 and their justice-related satisfaction were strongly correlated ($r(395) = .455$, $p < .001$). Participants' willingness to interact with their partner again correlated moderately with general affect at time 2 ($r(395) = .327$, $p < .001$), and strongly with participants' justice-related satisfaction ($r(395) = .431$, $p < .001$).

Effect of task allocation Participants could either assign one or two unpleasant tasks to their fellow partner as a reaction to their own unfair treatment. Participants' decision to either punish their partner by assigning two unpleasant tasks ($n = 209$ [53%]) or to behave fairly by assigning one fun and one unpleasant task to each ($n = 188$ [47%]) was equally distributed across conditions, $\chi^2(1, 397) = 3.82$, $p = .431$, and did not significantly affect participants' level of justice-related satisfaction, $F(1, 395) = 2.21$, $p = .138$. Neither did task allocation moderate the effect of experimental manipulation on justice-related satisfaction; interaction, $F(4, 387) = 1.64$, $p = .164$.

Discussion

Study 3 identified the expression of remorse as one marker of authenticity to which people pay attention when they receive feedback from a transgressor. If remorse was present, participants showed less rumination and more cognitive closure, as measured by the justice-related satisfaction scale. These findings highlight that people are sensitive to indicators of transgressor change on all three affective, behavioral, and cognitive levels. Participants were least satisfied if transgressors' feedback on change lacked in all of the three levels; were more satisfied the more indicators of change were

added and most satisfied if a transgressor showed indicators of change on all three levels. These findings support the hypothesis that transgressors need to demonstrate that their change is authentic and genuine and do not only need to pay lip service to change in order to make punishers satisfied.

Participants' general affect showed a similar pattern, suggesting that transgressor change also leaves people in a better mood than no change. Interestingly, participants were not equally sensitive to the authenticity of change when they filled out manipulation check items or the newly added variable that assessed participants' willingness to interact with their fellow Mturker again. For those items, it seemed that the presence or absence of transgressor remorse did not make a difference. Therefore, in regard to the manipulation check items, it seems that people do not need to be aware of the presence or absence of remorse in order for it to affect their justice satisfaction.

In regard to people's willingness to interact with fellow Mturker again, these findings further suggest that people can differentiate between their personal sense of psychological closure after a transgression, as indicated by the justice-related satisfaction scale, and their own responsibility to social functioning, as indicated by their willingness to interact with the other Mturker again. In some social groups, it is impossible to avoid any kind of contact with a person who has behaved unfairly. People who did not receive authentic feedback about change seem not to close the chapter; they are not satisfied, yet it seems they also know that they need to cope with their experience and manage to coexist with people they cannot ban from their community.

Interestingly, there were no differences in justice-related satisfaction between participants who decided to punish by allocating two boring tasks to their interaction partner and participants who decided not to punish by assigning one boring and fun tasks to each of them. There are at least two different ways to interpret the lack of differences. On the one hand, these results suggest that transformative justice motives generally arise after transgressions and are expressed through punishment or other means. On the other hand, however, one could critically argue that these findings suggest that justice-related satisfaction might not be suited to draw any conclusions on transformative justice motives because something that does not differ between punishers and nonpunishers cannot explain why punishers have decided to punish whereas nonpunishers have not. The present dissertation certainly supports the first interpretation. The design of Study 3 is not suited to completely rule out the second interpretation. Therefore, Study 4 continued to investigate different reactions to unfair behavior to further study potential differences between punishers and nonpunishers.

As a possible solution to the interpretation dilemma above, Study 4 aimed at studying potential origins of transformative justice motives. In order to better understand the mechanism of the effect of transgressor change on justice-related satisfaction, Study 4 investigated a potential moderating variable that determines if the effect of transgressor change on justice-related satisfaction is the same for punishers and nonpunishers or if it is different.

Study 4: The social origins of transformative justice motives

The goal of Study 4 was to investigate a potential origin for people's transformative justice motives. It was hypothesized that transformative motives have social origins because they ultimately seem to aim at a reintegration of the transgressor and to restore a social balance. Transformative justice motives are thus hypothesized to be most important in contexts when the likelihood of future interactions is high, in social ingroups.

Previous research suggests that it makes a difference to one's desire to punish if a transgressor is part of the person's social ingroup or outgroup. People have a strong need to see themselves and their ingroup in a positive light (Hogg & Abrams, 1988; Tajfel & Turner, 2001). Ingroup offenders pose a problem for such a positive perception as they threaten the validity of norms and shared values and call for a revalidation of these norms (Mendoza, Lane, & Amodio, 2014; Okimoto & Wenzel, 2009; Vidmar, 2002; Wenzel & Thielman, 2006). As a consequence, people are more likely to sanction the behavior of an ingroup transgressor than that of an outgroup transgressor (called the "black sheep effect"; Marques, Abrams, & Serodio, 2001; Marques, Yzerbyt, & Leyens, 1988; Shinada, Yamagishi, & Ohmura, 2008). Similarly, when people sanction a behavior, they expect an ingroup transgressor to feel more shame and embarrassment as reaction to the sanctioning behavior than an outgroup transgressor (Nugier, Chekroun, Pierre, & Niedenthal, 2009). And ingroup transgressors are judged more harshly and receive less forgiveness than outgroup transgressors when they do not show remorse (Gold & Weiner, 2000).

Whereas Studies 3 kept social category constant to the social ingroup of participant (i.e. the transgressor was either a fellow student or a fellow Mturker), Study 4 manipulated the social category of the transgressor in order to test whether it moderates the effect of transgressor change on justice-related satisfaction. It was hypothesized that people are particularly satisfied with transgressor change if the transgressor is part of their social ingroup. If the transgressor is part of the social outgroup, however, it was hypothesized that transformative justice motives they have social origins should be switched off, not leading to differences in justice-related satisfaction after the presence or absence of transgressor change.

In addition to only distinguishing between participants who punished their interaction partners and participants who treated them fairly (as in Study 3), Study 4 differentiated more within this latter group of participants by adding a third allocation option in which participants could decide to let the computer allocate the tasks randomly. Study 4 was therefore able to differentiate between participants who purposefully punished by assigning two boring tasks, participants who purposefully behaved fairly by splitting up the tasks evenly, and participants who simply preferred a fair allocation procedure by letting the computer decide randomly.

Lastly, Study 4 continued to explore individual difference measures to get a better understanding of the relationship between transformative justice motives and personality variables. In addition to participants' scores on victim sensitivity, self-control, and personal need for structure, Study 4 assessed participants' social identification with Princeton in order to explore if participants' shared identity with the transgressor moderates the effect of the experimental manipulation of ingroup

outgroup transgressors. Furthermore, the role of interdependent self-construal was explored. Previous research findings suggest that people with interdependent self-construal perceive an offense as more damaging to society and as a greater threat to shared values (Gollwitzer & Bueklein, 2007; Okimoto et al., 2012). It is possible that interdependent compared to independent participants have a stronger desire to effect a change in the transgressor through the imposition of punishment.

HYPOTHESIS 4:

Transformative justice motives are stronger in ingroup settings than in outgroup settings. The difference between the two feedback conditions (transgressor change versus no change) will be greater for ingroup transgressors than for outgroup transgressors.

Method

Study 4 was a laboratory study using a 2 (feedback: change versus no change) × 2 (social category of transgressor: ingroup versus outgroup) between-subjects design.

Procedure. Princeton undergraduate students ($N = 199$) were recruited from the psychology subject pool to participate in a study on "Office Behavior" and received half an hour of course credit for participation. When they arrived at the lab room, they were paired with an ostensible other participant who was five minutes late.

The ostensible other participant, one of four five experimental confederates who did not know the computer part of the study and therefore also did not know about the different feedback conditions, was randomly assigned to the social category condition prior to each individual session by using a mobile coin-flip application. In the ingroup condition, the confederate pretended to be another student and visibly carried a Princeton water bottle and a backpack. In the outgroup condition, the confederate pretended to be a "townie," i.e., a resident of the Princeton area unaffiliated with the

University, and did not bring these typical student accessories. After the confederate arrived, the experimenter checked the participant's and the confederate's IDs to adhere to the subject pool's guidelines. In the ingroup condition, confederates showed their Princeton student ID, which is what students usually use in these settings, visibly to the participants. In the outgroup condition, they showed their driver's license.

Then the experimenter left the room and pretended to quickly set up the two running rooms so that the participant and the confederate were alone in the lab space in a waiting room setting. All of the confederates had been trained individually to adhere to a standardized script to behave disinterested and somewhat bored regardless of the experimental condition, as well as to establish the ingroup or outgroup condition in the same way regardless of confederate identity. In the ingroup condition, confederates pretended to be undergraduate students during small talk (reporting their actual major and class year) who participate in the study for payment. In the outgroup condition, the confederates pretended not to go to Princeton but to live in a town nearby (Lawrenceville) and to work in Princeton at an office store (using a specific street name). They added that they sometimes participate in psych studies on campus to earn some extra cash. If participants asked about it, confederates pretended to be on a gap year between high school and college. In each session, confederates told the participants that they hope the study wouldn't take too long, which served as a prompt for the experimenter to return and guide the two participants to two separate, adjacent running rooms.

The rest of the study was administered by the computer. Participants were told that the study examined the extent to which performance can be affected by virt

environments and that they and the other participant would be asked to simultaneously work on office-related tasks. After entering a nickname for the study that would be shown to their partner, participants were asked to simultaneously send a business card. Participants would subsequently always receive a message from an invisible partner, who had apparently entered „hfjsdkghf“ as nickname, saying, „[nickname of participant], hope this won't take too long“.

Then the study procedure was explained, which was very similar to Study 3: Participants were told that the study would be about how they use the computer to work on several tasks related to office behavior. There would be different kinds of tasks, and they will have the opportunity to assign tasks to themselves and two tasks to their partners. Their partners would do the same. In total, they would each work on tasks that take about 20 minutes in total. First, „hfjsdkghf“ would choose two tasks to complete and assign two tasks to [nickname of participant]. Next, [nickname of participant] would choose two tasks to complete and assign two tasks to „hfjsdkghf“. After they decided who will work on which tasks, they would both work on their tasks simultaneously.

Before the task allocation started, participants filled out the same PANAS items used in Study 3 (Time 1, ... = .80). Next, participants watched how „hfjsdkghf“ was asked to either let the computer determine randomly how to allocate the first four tasks (consisting out of two fun and two boring tasks, taking about five minutes each) or to allocate them him-/herself. As operationalization of unfair treatment, „hfjsdkghf“ always decided to allocate the tasks him-/herself and subsequently assigned the two fun tasks to him-/herself („Watch a short clip of „The Office“ and answer some fun questions“ and „Check applicants' Facebook profiles and make a hiring decision (entertaining)“).

and the two boring tasks to the participant (,Highlight office-related words in a document, tedious but importantf and ,Check office supply list for typos, be sure not to miss onef).

Participants were asked to indicate how much they are looking forward to do the tasks that ,hfsdkghf has picked for them (1=not at all, 7 =very much). Next, as operationalization of their reaction to unfair treatment, participants were asked if they wanted to either let the computer determine randomly how to allocate the remaining four tasks (again consisting of two fun and two boring tasks, taking about 5 minutes each) or to allocate them themselves. If participants decided to let the computer determine randomly, they would see that the computer gave one boring and one fun task to each of them. If participants decided to allocate them themselves, they would drag and drop the two fun tasks (,Rate comics related to general office behavior, funf and ,Rate comics related to social dynamics in an office, entertainingf) and the two boring tasks (,Count numbers of letters sent out last week, tedious but importantf and ,Check copy room supply list for typos, be sure not to miss onef) to the name of the person to which they wanted to allocate them.

Participants subsequently read that the tasks now have been allocated, and that before both participants would start working on their assigned tasks, the two of them could now simultaneously send one last message to each other. Participants, the following page where they would enter their message only appeared for three seconds and then automatically advanced, showing the error message ,Page error. Message could not be sentf, followed by a waiting page that indicated that the participants were still writing a message to them. In the changed conditions it read ,got it, you think that wasn't very

nice of me.. but sometimes i gotta be selfish.. i would do the same thing, again in the change conditions it read ,got it, that wasn't very nice of me.. shouldn't have been so selfish.. i would do it differently now if i could

On the next page, before the tasks would start, participants were asked to answer questions about the interaction with ,hfjsdkghf using the same justice satisfaction scale that has been used in the previous studies (12 items), followed by the question ,Would you like to interact with ,hfjsdkghf again in the future? (1 = not at all, 7 =very much).

In addition participants answered the same manipulation checks that were used in Study 3... = .76) and were asked to indicate in an open ended text field anything they remembered about their interaction partner ,hfjsdkghf. These comments were coded as to whether they included information about the partner's social category

At the end of the interactive part of the experiment, participants indicated how they feel right now (PANAS time 2 ... = .82) They then worked on Task 1 and rated comics on how funny they were, as well as on task 2 where they checked a list for typos. For the sake of time, Tasks 3 and 4 apparently could not be done. Next, participants indicated once again how they feel (PANAS time 3 ... = .86)

Subsequently participants were asked to fill out some individual difference measures. In order to assess students' social identification with Princeton, five items were taken from an ingroup identification scale (Leach et al., 2008) and adjusted to the Princeton student context. = .82) e.g., Being a Princeton undergraduate student is an important part of how I see myself. As a second individual difference measure, participants' self-construal was assessed using the 16-item scale by Gollwitzer and

Bücker (2007) that consists of three subscales: subscale interdependent/ context dependence (... = .77) e.g., In general, other people's behavior influences my own behavior; subscale interdependent/ relatedness (... = .79) e.g., It is always very important to me that the people who are close to me are feeling good; subscale independence (... = .76) e.g., I like being unique and in many respects different from other people. In addition, the same scales that were used in Study 2 assessed participants' level of victim sensitivity (... = .91), sense of control (... = .64), and personal need for structure (... = .85). See Appendix for a list of all items.

At the end, participants indicated their age, gender, political orientation, and ethnic background. Using the same continuous suspiciousness items as in Study 3, participants indicated how much they doubted that the other partner was real and how much it affected their behavior. Participants were debriefed about the deception aspects of the study and gave informed consent that their data can be used. At the very end, participants were asked if they had heard about the true purpose from other students, and if they knew from other people that the partner was real (yes/no).

Results

Preliminary analyses In 25 of the 199 cases, the experimental manipulation of the transgressor's social category (townie or student) could not be established, for instance because participants knew the confederate they were paired with or because participants did not react to the confederate's attempts to make small talk. Additionally, participants who did not follow the instructions when they were asked to assign tasks (e.g., who did not assign any tasks at all and only clicked next or only assigned tasks to themselves but not to the other person 9), participants who purposefully allocated

two boring tasks to themselves ($n = 7$), and participants who indicated that they knew from other students that the partner was not real ($n = 3$) were excluded from the analyses, leaving a final sample $N = 156$ ($M_{\text{age}} = 19.5$, $SD_{\text{age}} = 1.1$; 92 [59%] female participants equally distributed across conditions) with 42 participants in each experimental condition. All of the findings discussed below remained significant if the analyses were conducted with the complete dataset ($N = 159$).

There were several continuous items asking about participants' level of suspiciousness about the setup of the study, one of which differed between experimental conditions, and none of which correlated with the main dependent variable justice-related satisfaction, $r(154) < |.092|$, $p \leq .253$. Participants who indicated they were suspicious that the study was only about office tasks were equally distributed across conditions, $\chi^2(3, N = 156) = 1.54$, $p = .673$, and this suspicion did not correlate with the main dependent variable, $r(154) = -.077$, $p = .341$. None of the participants indicated to have heard about the true purpose from other people before. Lastly, there was an open-ended comment field that was used to manually code the presence of participants' suspiciousness about the truthfulness of the received messages about the confederate status of the participants' interaction partner ($n = 30$). This latter indicator of suspiciousness was not equally distributed across the experimental conditions, $\chi^2(1, N = 156) = 8.48$, $p = .037$, reflecting that 14 of the 30 suspicious participants were in the outgroup/no change condition. However, the presence of this kind of suspiciousness did not correlate with the main dependent variable, $r(154) = -.045$, $p = .580$. All of the significant findings presented would remain significant if the analyses controlled for the continuous suspiciousness measures. If the 30 suspicious participants were excluded, the

pattern of results would not change. Therefore, these suspicious participants remained in the final dataset.

Each of the five different confederates (four female, one male) who either pretended to be students, townies or participating in the study for money performed in 20 to 42 different experimental sessions. The confederates were equally distributed across experimental conditions ($F(12, N = 156) = 7.99, p = .786$). Confederate identity did not have any effect on the main dependent variable justice satisfaction, $F(4, 151) = 0.95, p = .436$. Similarly, the confederates' gender (coded as female = 1, male = 0) had no effect on the dependent variable ($F(1, 154) = .027, p = .742$), and there were no interaction effects of the participant's gender and the actor's gender on the dependent variable, $F(1, 150) = 0.37, p = .542$. Therefore, all of the analyses collapsed across confederate identity. Participant gender also did not correlate with the dependent variable, $r(152) = -.09, p = .265$.

Manipulation checks

Perceived change in the transgressor. Two-way ANOVA results confirmed that the experimental manipulation of transgressor change was successful, $F(1, 152) = 211.91, p < .001, \eta^2 = .58$, showing higher ratings in the change conditions for both the in-group and out-group conditions (in-group: $M = 5.01, SD = 1.09$; out-group: $M = 4.79, SD = 0.99$) than in the no change conditions (in-group: $M = 2.59, SD = 1.10$; out-group: $M = 2.42, SD = 0.92$). Thus, participants in the change conditions agreed more than participants in the no change conditions that their task allocation has caused a change in their partner's attitude, that their partner regrets her/his allocation of tasks, and that their partner would probably not act the same way again in the future.

manipulation of the transgressor's social categorization did not affect the manipulation check on transgressor change ($F(1, 152) = 1.42, p = .236$), neither did the interaction of social categorization and presence or absence of change ($F(1, 152) = 0.02, p = .887$).

Perceived social category of the partner. Participants' open-ended responses about what they remembered about their interaction partner were coded for outgroup conditions in regard to whether participants mentioned that the other person was not a student but someone who works in town, assuming that the default situation students encounter is one in which another participant is a student like them. Only the 80 (45%) participants in the outgroup conditions explicitly mentioned this detail about their interaction partner. Compared to the ingroup condition, this rate is satisfactory since only 15 of the 76 (20%) participants in the ingroup conditions explicitly mentioned that their interaction partner was a fellow student, which is in line with the assumption about a fellow participant's student status by default. However, 55% who fail to mention spontaneously that their interaction partner is a townie are a too substantial amount of participants to exclude from the final dataset in order to keep statistical power for the planned analyses. The perceived categorization of the partner mentioned in the open-ended responses was therefore not used to exclude any participants. Nevertheless, if subjects who failed to mention that their partner was not a student were excluded from the outgroup conditions, the patterns of results could be stronger (see results section on justice-related satisfaction for details).

Effects of transgressor change on general affect, justice-related satisfaction, and willingness to interact again

General affect. PANA scores were measured three times in Study 4, allowing the assessment of temporal dynamics in general affect. A significant main effect within subjects indicated that participants were generally less satisfied at time 2 when they had just finished the interactive part with their ostensible partner ($M = 4.65, SD = 1.04$) compared to the very beginning of the experiment at time 1 ($M = 5.36, SD = 0.93$), and that participants were then again more satisfied at the end of the experiment, at time 3 ($M = 4.99, SD = 1.09$) PANA $F(2, 304) = 40.44, p < .001, \eta^2 = .210$. There were no significant interaction effects of the repeated PANA measures with the experimental conditions, $.146, ps < .546$. Similarly, there was no main effect of the experimental conditions on the general level of positive or negative affect; change versus no change $F(1, 152) = 2.26, p = .134$ ingroup versus outgroup $F(1, 152) = 0.00, p = .975$; interaction term $F(1, 152) = 0.10, p = .751$.

Justice-related satisfaction. If the effect found in previous studies has social origins, a change in the transgressor's attitude should matter more for a transgressor who belongs to the participant's ingroup than for a transgressor who is part of the participant's outgroup. Although two-way ANOVA results revealed a significant main effect for change manipulation $F(1, 152) = 5.33, p = .022, \eta^2 = .034$, they did neither reveal a main effect for ingroup/ outgroup manipulation $F(1, 152) = 0.31, p = .576$, nor a significant interaction effect, $F(1, 152) = 0.49, p = .485$, possibly caused by a lack of statistical power. Nevertheless, direct contrast analyses revealed that the pattern of results was in support of the hypothesis. Participants in the ingroup conditions were

significantly more satisfied when the transgressor showed a change in attitude ($M = 4.76, SD = 0.71$) than when the transgressor showed no change in attitude (change $M = 4.40, SD = 0.67$), $t(74) = 2.28, p = .026, d = 0.52$. As predicted, the difference between the change and no change feedbacks for outgroup conditions on the other hand, was smaller and not significant (change $M = 4.63, SD = 0.87$; no change $M = 4.41, SD = 0.72$), $t(78) = 1.08, p = .284, d = 0.24$. If one only looked at the participants in the outgroup conditions who spontaneously mentioned in their open ended comments that their fellow interaction partner was someone unaffiliated with the University, the difference between the change ($n = 18$) and no change condition ($n = 18$) would become even smaller (change $M = 4.67, SD = 0.97$; no change $M = 4.74, SD = 0.58$), Welch's $t(27.9) < 1, p = .796, d = 0.09$.

Willingness to interact again Replicating previous studies, participants in the change conditions ($M = 4.37, SD = 1.34$) agreed more strongly that they would like to interact with their ostensible interaction partner again in the future than participants in the no change conditions ($M = 3.63, SD = 1.52$), $F(1,152) = 10.60, p = .001, \eta^2_p = .065$. Whether the transgressor belonged to the participant's ingroup or outgroup did neither have a main effect, $F(1,152) = 0.15, p = .697$, nor an interaction effect with the presence or absence of transgressor change on the participants' willingness to interact with their partner again in the future $F(1,152) = 0.41, p = .523$.

Intercorrelations of the dependent variables General affect at time 2 and justice-related satisfaction were strongly correlated ($r(154) = .515, p < .001$).

Participants' willingness to interact with their partner again correlated moderately with

general affect at time 2 ($t(154) = .341, p < .001$), as well as moderately with participants' justice-related satisfaction ($t(154) = .354, p < .001$).

Effects of task allocation. As described above, participants could either decide to let the computer determine randomly who would work on which tasks ($n = 37$), or could decide to allocate two fun tasks to themselves and two boring tasks to their partner, respectively ($n = 65$), or they could purposefully assign one boring and one fun task to each their partner and themselves ($n = 54$). Task allocation was distributed equally across all of the conditions, $F(1, 156) = 4.55, p = .034$, including across the manipulation of social categorization ($F(2, N = 156) = 0.98, p = .378$), reflecting that participants decided equally often to punish ingroup and outgroup transgressors.

In regards to justice-related satisfaction, if type of allocation is added as an additional factor to the ANOVA, it does not significantly predict justice-related satisfaction ($F(1, 144) = 0.58, p = .559$). Nevertheless, the descriptive pattern of results reveals some important observations (sample size in each cell = 15). It seems that the difference between ingroup and outgroup transgressors for the different types of task allocators. If participants decided to punish their partner by allocating boring tasks, change in an outgroup transgressor led to the same increase in justice-related satisfaction as change in an ingroup transgressor. Interestingly, the difference between ingroup and outgroup transgressors was particularly distinctive if participants decided to be fair in particular way: not by letting the computer assign randomly but by purposefully allocating fairly, that is one boring and one fun task to each.

Figure 5

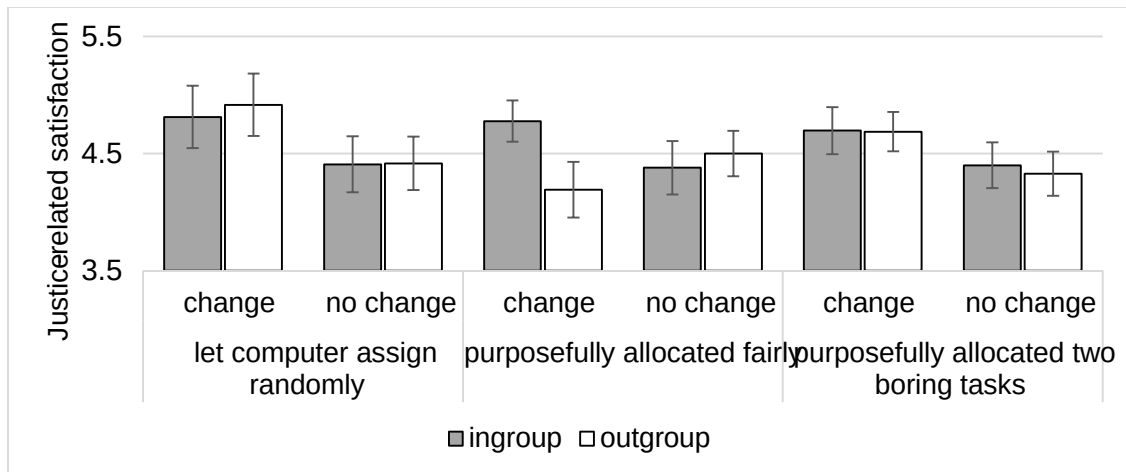


Figure 5. The effect of transgressor change on justice-related satisfaction in Study 4 seems to vary for different types of task allocators (punishers and different kinds of nonpunishers)

Effects of individual difference measures Victim sensitivity was the only measure that significantly varied between different types of task allocators, $F(2, 81.4) = 5.25, p = .007, \eta^2 = .070$, indicating that participants who decided to purposefully allocate two boring tasks to their partner scored higher on victim sensitivity ($M = 4.95, SD = 1.00$) than participants who purposefully allocated the tasks fairly ($M = 4.51, SD = 1.15$) and participants who let the computer assign randomly ($M = 4.13, SD = 1.53$).

Only victim sensitivity, $r(154) = -.171, p = .033$, and the self-construal subscale interdependence/ relatedness correlated significantly with justice-related satisfaction, $r(154) = .208, p = .009$, indicating that people who score low on victim sensitivity and people who score high on interdependence/ relatedness are generally more satisfied. Justice-related satisfaction did not correlate with any of the other measures assessed, $.343 < p < .805$.

Interaction analyses using multiple regression revealed no significant differential effects for any of the entered individual difference measures in combination with one or both of the dummy-coded experimental manipulations on the main outcome, except for the variable social identification with Princeton which showed a significant three way interaction with the two experimental manipulations, $F(7, 148) = 3.77, p = .040$. (Main effects for social identification $\beta = .162, p = .331$; transgressor change $\beta = .231, p = .044$; transgressor's social category $\beta = .020, p = .861$. Interaction terms for social identification*transgressor's social category $\beta = -.264, p = .136$; social identification*transgressor change $\beta = -.192, p = .268$; transgressor's social category*transgressor change $\beta = -.100, p = .468$. Model $F(7, 148) = 15.8, p = .145, R^2 = .070$.)

This interaction describes the surprising finding that participants with low social identification with Princeton show the predicted effect of transgressor change on justice related satisfaction only for ingroup transgressors but not for outgroup transgressors, whereas participants with high social identification with Princeton show the predicted effect for both ingroup and outgroup transgressors. Figure 6 illustrating the meaning of the interaction term using mediaplots.

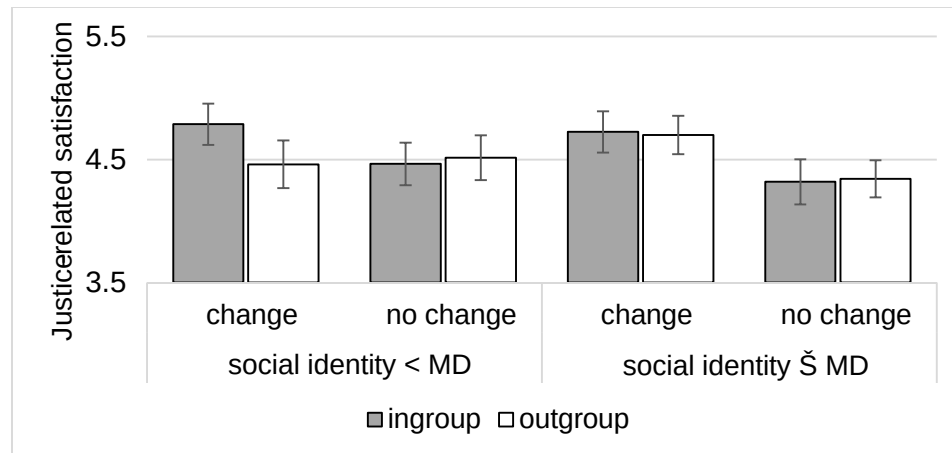


Figure 6. The effects of the transgressor's social categorization and transgressor change on justice-related satisfaction in Study 4 depicted separately for participants with social identification scores below and above the median

Discussion

Results from Study 4 suggest that transformative justice motives have social origins. Replicating effects found in previous studies, participants were more satisfied after transgressor change than after no change. However, this difference was only significant for ingroup transgressors and not for outgroup transgressors.

There are two important aspects about this finding. First, it demonstrates that transformative justice motives are particularly strong in ingroup contexts. Although the ANOVA interaction term did not reach significance, presumably due to statistical power issues, contrasts testing the predictions directly were significant. Moreover, if only those participants remained in the analyses who correctly mentioned in their comments that their interaction partner was not affiliated with the University, the pattern of differential results was even stronger, supporting the interpretation that transformative justice motives predominantly arise in ingroup settings.

Second, these differential findings highlight that a change in the transgressor does not always lead to higher justice-related satisfaction. By demonstrating that justice-related satisfaction is sensitive to the transgressor's social category, results from Study 4 suggest that such satisfaction can be interpreted as indicator for the fulfillment of certain transformative goals (see Discussion Study 3).

One important limitation of these differential findings for ingroup and outgroup transgressors refers to the extent to which the decision to punish or not (participants' task allocation) presumably affected people's justice-related satisfaction after the presence or absence of a transgressor change. Although not significant, there were different patterns of results for different types of task allocation. Participants who purposefully allocated the tasks fairly (i.e. participants who did not punish) seemed to have acted on transformative justice motives toward ingroup transgressors, but not toward outgroup transgressors. Participants who purposefully assigned two boring tasks to their partner (i.e. participants who expressed their transformative justice motives by punishing) however, did not seem to differentiate between ingroup and outgroup members; these punishers were more satisfied after transgressor change than after no change regardless of the social category of the transgressor. These findings suggest that in ingroup contexts transformative justice motives are not exclusive to punishment behavior; reacting to a transgression with purposefully fair behavior can also serve transformative functions and can be used to send a message to the transgressor or demonstrate how right behavior looks. Thus, transformative motives seem to have social origins as hypothesized; they generally arise after ingroup transgressions and can be expressed by punishment or other means. However, these findings also suggest that

transformative motives are not completely switched off for outgroup transgressors as originally hypothesized.

There are several open questions that remain to be investigated: Why did transformative justice motives spillover to the outgroup for some participants, and why are transformative justice motives expressed sometimes punitively and sometimes constructively toward ingroup transgressors? Study 4 continued to explore the role of personality measures in this context and how they affected people's justice-related satisfaction. Unlike in Study 2, there were two measures that correlated with justice-related satisfaction, one of which was victim sensitivity that was also related to participants' task allocations, like in Study 2. Victim sensitive individuals were more likely to punish transgressors yet showed lower ratings on justice-related satisfaction. Thus, victim sensitivity seems to be an important individual difference measure that should be explored further in future studies.

There was only one significant interaction found in Study 4 (the effect of personal need for structure found in Study 2 could not be replicated), and more surprisingly, this three-way interaction of social identification with Princeton with the two experimental manipulations showed a somewhat surprising effect: participants who identified less with Princeton showed the expected ingroup/outgroup difference but participants who identified more with Princeton did not. At first sight, this effect looks similar to the difference between participants who purposefully allocated the tasks fairly and people who purposefully allocated two boring tasks to their partner. However, social identification with Princeton did not significantly vary between the different types of task allocators, $F(2, 153) = 1.52, p = .221$; and low and high identifiers (medians split)

were equally distributed across the three types of task allocators $F(1,26) = 1.38$, $p = .501$. Similarly, low and high identifiers did not differ in any of the suspiciousness measures.

All in all, the interaction findings is counterintuitive at first sight, and admittedly, it is a small effect in terms of effect size ($r^2 = .07$, and the full model is not even significant, see Results section). Although it could just be a false positive finding and needs to be replicated in future studies, this finding might point toward an interesting hypothesis that remains to be tested: it is possible that the social origin of transformative justice motives is not so much due to identification with one's ingroup but due to the increased likelihood of future interactions. If people know that they have to get along well, they might think to better act prosocially toward each other and try to reintegrate transgressors to restore their togetherness. Thus, a difference between the treatment of ingroup and outgroup transgressors would ultimately be caused by a different perception of the likelihood of future interaction and not by the desire to see the ingroup in a positive light (Hogg & Abrams, 1988; Tajfel & Turner, 2001). In the special case of Study 4 outgroup members who come to campus could be perceived in more or less elaborate ways depending on the participant's social identification with Princeton. For participants who value the distinction between ingroup and outgroup by identifying strongly with Princeton, townies on campus raise the perceived likelihood of future interactions with the outgroup and might increase participants' transformative justice motives toward outgroup individuals. Strongly-identifying Princeton students might want to educate students and non-students alike about how one is supposed to behave on campus.

On the other hand, students who do not really identify that much with Princeton do not pay as much attention to the presence of outgroup members and just think that future interactions with outgroup members are less likely, leading to the predicted difference between ingroup and outgroup transgressors. A difference that might then not be caused by the transgressor's social categorization but by the perceived likelihood of future interactions. Since the direction of the interaction effect was not predicted, more studies are needed that vary the environment in which the transgression occurs in order to disentangle the effect of social category and the perceived likelihood to interact again. Future studies need to examine if the difference between effects for ingroup and outgroup transgressors is mostly caused by social categorization or by the likelihood of future interaction. It is possible that the dynamics found in this study differ for contexts where the environment is neutral and the outgroup member is not coming to the ingroup's territory and therefore does not increase the perceived likelihood of future interactions for some participants.

In Study 4, general affect did not show the same differential pattern as justice related satisfaction, suggesting once again that the fulfillment of transformative justice motives is not necessarily reflected in general mood. It was the first time in Study 4 that general affect was measured at three time points, compared to two time points in previous studies. Adding another time point revealed that general affect first decreased but then seemed to recover at time 3, at the end of the experiment. It would be interesting for future studies to also assess justice related satisfaction at more than one time point in order to explore its change or stability over time. To get a better understanding of the temporal dynamics of justice related satisfaction, it would be

particularly interesting to study if justice-related satisfaction still differs between experimental conditions at the very end of the study. If transformative justice motives are strong, one would expect that the differences between change and no change conditions persist.

Participants' willingness to interact with their partner again in the future was affected by the presence versus absence of transgressor, replicating the pattern of results of Study 3. Unlike justice-related satisfaction, the transgressor's social categorization did not affect participants' willingness for future interactions. Like in Study 3, this finding suggests that people potentially differentiate between how they feel about a transgression and how they feel about behaving toward the transgressor. Even if people might not really have closed the chapter, their willingness to interact with the transgressor again reflects potential for reconciliation. Future research could try to unfold the temporal dynamics of how people's willingness to interact with a past transgressor might affect their justice-related satisfaction. It is possible that such an early willingness to interact again might actually help people to reach their fullest potential of reconciliation and ultimately increase justice-related satisfaction over time.

One limitation of Study 4 was that the intended manipulation check for the manipulation of the transgressor's social category did not necessarily identify all of participants who remembered whether their interaction partner was affiliated with the University or not. The results strongly suggest that the manipulation was successful, as do the findings for those participants who correctly indicated in their comments that their interaction partner was not a student. Nevertheless, future studies should include items that better verify if the manipulation of social category was successful.

studies should also aim for a paradigm that replicates the effect on ~~justified~~ satisfaction with a bigger sample so that ANOVA interaction terms ~~can~~ become significant.

In Study 4, the frequency of punishment did not differ between ingroup transgressors and outgroup transgressors. Many participants decided to refrain from punitive responses towards ingroup transgressors ~~to use~~ more constructive means in order to reinforce the violated group norm (Okimoto & Wenzel, 2009). Other studies have found that ingroup transgressors are more likely to be punished than outgroup transgressors (Marques et al., 2001; Marques et al., 1988; Shinada et al., 2004). Nevertheless, the current findings are not necessarily in ~~opposite~~ findings on the , black sheep effect. In order for increased ingroup harshness to occur, guilt needs to be certain (van Prooijen, 2006) or needs to be inferable from repeated transgressions (Gollwitzer & Keller, 2010), otherwise ingroup transgressors receive the benefit of the doubt. It is possible that the paradigm used in Study 4 was ~~suited~~ to detect increased harshness against ingroup transgressors, for instance, because the transgression was not severe and did not seriously enough question ingroup norms. ~~from the~~ , black sheep effect / to emerge

Future studies should try to differentiate between transformative justice motives (i.e. desire for transgressor change) and sensitivity toward cues of change (i.e. the detection of change in ingroup versus outgroup transgressors). Although participants in Study 4 were equally capable of detecting change in ingroup and outgroup transgressors as indicated by the manipulation check, other research findings suggest that people are more likely to attribute complex emotions such as remorse to ingroup members than to

outgroup member (Wohl, Hornsey, & Bennett, 2012), a phenomenon that has been linked to the inhumanization of outgroup members (Leysens et al., 2001) and that makes intergroup reconciliation challenging. In the context of punishment severity, Lieberman and Link (2007) found that outgroup offenders receive lower remorse ratings and are punished more harshly than ingroup offenders for the same offense (yet for both groups the offenses received the same ratings on moral wrongness). Although these findings support a notion of transformative justice motives and their absence in ingroup contexts, it remains to be examined under what circumstances people actually aim less at a change in outgroup transgressors versus when people do not detect a change in outgroup transgressors even if it occurs. Studies that manipulate the transgressor's social category and display of remorse remain an interesting avenue for future research.

General Discussion

Using different experimental stimuli and procedures, the present set of studies provides empirical evidence for people's transformative justice motives, aiming at genuine change in a transgressor's attitude and behavior. Previous research has shown that punishers and reprimanders were more satisfied if transgressors changed their attitudes and understood that what they did was wrong. This effect was enhanced the more authentic a transgressor's change appeared to be, indicated by the presence of remorse. Moreover, the increase in justice-related satisfaction after transgressor change could not be explained by inferred suffering.

All in all, the present set of studies suggest that retributive punishment theories have a much more limited applicability than previous research has suggested.

Retribution still has a role in punishment but not the major one. People do not only seem to give transgressors what they deserve. They hope to transform transgressors and make them understand that what they did was wrong. As social beings, people seem to strive for reconciliation and hope to achieve a change in attitude in those who have treated them unfairly. In other words, punishment can only make people change their way, if punishment is inflicted on the people who deserve it. If punishment is retributive, see Moberly, 1968. The underlying motive and ultimate goal for such punishment, however, does not seem to be retributive but transformative.

Transformative justice motives can explain why people prefer to punish proportionally. Psychologically, weighing out the scales of justice does not refer to simply paying back and giving transgressors what they deserve. It appears to follow a hydraulic principle that carefully calibrates blame and guilt (Malle, Monroe, & Guglielmo, 2014). If a person commits a transgression and creates an imbalance of justice, the imbalance gets first addressed by blame that relates to the degree of guilt a transgressor experiences. The less guilt transgressors experience the more people blame them, and the more transgressors experience guilt the less people blame them (Chenbo & Watanabe, 2009; Scher & Darley, 1997). For the transgressor, the experience of guilt is a first step to accept responsibility and to make amends to the victim (Fisher & Exline, 2006; Houser, Windzio, & Greve, 2008) and thereby clears the way for reconciliation. The experience of guilt signals that the transgressor cares and will behave differently in the future (Baumeister et al., 1994). People are very good at coding if the expression of guilt (i.e. remorse) is real or fake (ten Brinke, MacDonald, Porter, & O'Connor, 2012). Importantly, if people see that transgressors feel genuinely bad about what they did (i.e.

feel guilty or remorseful) and consequently blame them less, they also show less of a demand for harsh punishment because the imbalance of justice has been partly addressed (Bornstein, Rung, & Miller, 2002; Darby & Schlenker, 1989; Pipes & Alessi, 1999; Robinson et al., 1994; Rumsey, 1976). If transgressors don't feel bad about what they did, however, and the scales of injustice are not evened out, punishment can be used as a guilt-inducing strategy (Petersen, Sell, Tooby, & Cosmides, 2010, p. 101, also see communicative theories on punishment).

The present dissertation suggests that the foundation for this hydraulic restoration of justice through the expression of blame and punishment on the one side and the acceptance of guilt and remorse on the other lies in people's transformative justice motives. These transformative justice motives do not only refer to punishment behavior but ultimately also allow for reconciliation: Only if transgressors acknowledge that what they did was wrong and change their attitude, and ultimately their behavior, can they be reintegrated into a community or society (Bazemore, 1998) enabling intragroup reconciliation and relationship repair (Okamoto & Wenzel, 2014).

At the same time, transformative justice motives can explain why punishment is sometimes not considered to be necessary. The present dissertation suggests in order to make transgressors experience guilt and remorse and to make them acknowledge that the transgression was unjust, some people decide to punish, but others do not. Non-punishers who purposefully demonstrate fair behavior as a reaction to unfair treatment use a smart way to effect transformation in a transgressor: transgressors who feel punished and ostracized, for instance, actually have less of a desire to reconcile with the victim to begin with (Woodyatt & Wenzel, 2013). Transgressors sometimes can

be more likely to change if victims signal potential forgiveness (Wallace, Exline, & Baumeister, 2008). Ironically, understanding people's punishment motivations as transformative might help to refrain from punishment in the first place and to find other, more effective ways to achieve people's underlying goals.

The present findings suggest an approach to punishment that focuses on person centered cues and on information about moral character instead of on features related to an action (see Goodwin, Piazza, & Rozin, 2014, demonstrating that information on moral character predominate in person perception and evaluation; Uhlmann, Pizarro, & Diermeier, 2015, for a review on a person centered approach to moral judgments). People's justice related satisfaction seems to be linked to how they evaluate the transgressor's character: if transgressors indicate through remorse and a change in attitude that, deep down, they are good people and understand what is right and what is wrong, people might either infer that the transgression was just a short episode, an atypical event for a person with otherwise decent character, or similarly they might infer that the person has now truly mended their ways and will not return to their past lifestyle. Thus, people seem to be virtue ethicists who focus not only on the badness of actions or the prevention of bad actions, but also on the goodness of character and how good character traits can be elicited in others.

There are several limitations to these findings and aspects of transformative justice motives that future studies need to address. Due to feasibility, in all of the presented studies transgression seriousness was quite low: unfairness was operationalized by small sums of money or boring short tasks. It is possible that transformative justice motives are affected by the seriousness of a transgression and that

more serious, real-world transgressions that potentially also have offensive, humiliating, or violent components evoke transformative justice motives to a different degree.

The present set of studies only examined intentional harms which are known to elicit more blame than unintentional harm (e.g., Ames & Fiske, 2013, 2015). Future studies will show if transformative justice motives also play a role for unintentional harms. Transformative justice motives are presumably less likely to occur after completely unintentional and unforeseeable harms. If an unintentional harm would have been controllable by the transgressor, however, it is possible that people hope to effect a change in the transgressor's carelessness (and to make the transgressor exercise greater care; see for instance Martin & Cushman, in press) thus also have transformative justice motives.

Moreover, in order to study the importance of genuine change versus mere behavior change, the present set of studies relied on feedback about intended behavior change that was either paired with a change in attitude and did not manipulate actual behavior change. Future studies on transformative justice motives should ideally include real behavior change and not just feedback about intended behavior change. Although behavioral studies have not directly addressed the question yet, there are some promising results coming from economic games research that potentially support the notion of transformative justice motives. Recent results from two-round trust games revealed that participants who had punished a previously uncooperative partner in the first round were more likely to cooperate with their partner in the second round than participants who had decided not to punish their cooperative partner, suggesting that punishers expect a transgressor to learn from punishment. Punishers were even as

likely to cooperate with a previously uncooperative partner as non-punishers who had interacted with a cooperative partner before (Krauss, Cosmides, Pedersen, & Tooby, 2012). From a transformative justice perspective, it is reasonable for punishers to cooperate with transgressors again after punishment because punishers in these situations presumably infer that their partner has truly changed their attitudes as a consequence of punishment and understand that their uncooperative behavior was wrong, and because punishers do not simply think that uncooperative partners merely show a change in behavior to avoid punishment.

It remains to be tested if the different findings in Studies 4 for ingroup and outgroup transgressors were caused by social category or different perceived likelihoods of future interaction, thus if transformative justice motives generally transcend group categories once people want to punish or not. At this point, the present findings suggest that transformative justice motives are particularly strong in ingroup context, as they are expressed through punishment and other means, but that they are probably also important for people who decide to punish outgroup transgressors.

If transformative justice motives have social origins in that they are linked to perceived likelihood of future interactions but not so much to the social category of a transgressor, these findings have important implications for the legal system. Although, in real-world cases, victims and transgressors do not necessarily share the same social category, future interactions between the victim and the transgressor are possible nonetheless.

Implications for the legal system

Transformative justice motives are theoretically better suited than retributive motives to explain the acceptance of current legal practices. First, they can help to explain the intuitive appeal of parole, which is the provisional release of prisoners before they have completed their sentence, when offenders have demonstrated that they have changed and that they can hope for a premature release from prison. For people who were retributivists, they should be against such practices. Second, a transformative approach to punishment is also better suited to explain why the presence of remorse is accepted as a mitigating circumstance at court (see Proeve & Tudor, 2010, Appendix Chapter 6, for an overview on how state and federal laws in the U.S. deal with remorse, as well as how remorse is dealt with in other countries' legal systems). Such remorse neither makes the past transgression less severe nor reduces the deservingness of punishment from a retributive perspective. Instead, such remorse indicates that a change has already partly occurred or is likely to occur in the near future. As a consequence, in line with transformative justice motives, less punishment is needed to evoke such a change.

Nevertheless, transformative justice motives also seem to have problematic aspects and would pose challenges to the concrete setting of sentences. If transgressors did not change, participants in the present studies did not agree that everybody got what they deserve. One implication of this finding is that if people do not have a sense that justice is restored in the absence of transgressor change, they might demand more or harsher punishment although the transgressor has already been punished. Thus, if used as a justification for punishment, relying on transformative motives when setting a

concrete punishment (i.e. punishing until a transgressor changes) could add to a ratcheting up effect of punishment severity (McGeer, 2013) that lets a penalty get out of proportion and ultimately violates people's intuition that punishment should fit the crime. Moreover, it would take away the transgressor's free will of choosing for him or herself how to react to punishment. The ideal punishment, although transformative in its scope, therefore needs to have an upper limit of punishment severity, for instance by relying on punishment that is proportional to a transgression's seriousness, like retributive or communicative punishment theories would suggest, e.g., Duff, 2001.

The present findings further suggest that in order to reach psychological closure, victims should get to know if a transgressor has changed, ideally not just in restorative justice settings. Actual crime victims are not satisfied after the mere imposition of punishment on a transgressor (Orth, 2004). Transformative justice motives suggest that victims need a specific reaction from the transgressor after punishment to experience satisfaction. Yet in contemporary justice settings, the opportunity for such communicative exchanges between offenders and victims is extremely limited. The current results suggest that, supposing punishment has effected a change in the transgressor's attitude, it might be psychologically beneficial for victims to hear this. At the very least, victims or their families should not be led to believe that their feeling of injustice can be attenuated by punishment alone.

More research is needed to test which stages of change are potentially more important for punishers or victims to perceive than others. Change in the real world does not only constitute a distinct event of transgressor feedback like in a psychological

experiment but is a process that consists of several stages from an initial intention to change to the repeated demonstration of different behavior.

Lastly, psychologists should investigate how to best help victims to achieve closure after the experience of injustice if transgressors do not change as a consequence of punishment. These findings could also be used to assist victims who decided to refrain from punishing the wrongdoer to reclaim control over their lives (e.g., rape victims who want to see the offender admitting the wrong but who do not want to take the rapist to court).

Next steps for psychological research on transformative justice motives

To better understand the generalizability of transformative justice motives, future studies should continue to explore people's transformative justice motives after transgressions by using different experimental paradigms that include different kinds of transgressions that vary in seriousness, different kinds of punishment, and that use different dependent variables. Only Study 2 used a paradigm where the transgressor's unfair behavior could not be evened out. Future studies should continue to use operationalizations that are at least related to compensation or a restoration of equity as possible, so that the operationalization of punishment is close to a pure punitive reaction. Possible sanctioning reactions to unfair treatment could include avoidance or exclusion behavior to broaden the range of possible punitive means for transformation. Non-punitive, constructive means for transformation could be operationalized and measured in the forms of compensation, direct interaction, or joint activities. In order to get a sense of how many participants actually punish (i.e. treat their interaction partner in a bad way as a response to what they did), it would also be interesting to assess a base rate of

the behavior that is interpreted as punitive reaction. Studies 3 and 4, for instance, it would have been interesting to know how many people would have assigned two boring tasks if they would have been first to assign tasks. Moreover, the transgressions used in future studies do not need to be based on deception but could make use of real-world interactions, for instance, by using ecological momentary assessments (see, e.g., Hofmann, Wisneski, Brandt, & Skitka, 2013; Shiffman, Stone, & Hufford, 2008).

In addition, future studies should try to identify the factors that determine whether people punish or not, examine the role of spontaneous change, as well as study potential mediators for the effect of transgressor change on justice-related satisfaction.

What are the factors that determine if transformative motives are expressed through punitive means or constructive means? Many subjects in the present set of studies decided not to punish the transgressor. None of the assessed individual difference measures was able to really explain the distinction between punishers and non-punishers. Future research is needed to identify the factors that make people pursue or refrain from punishment. These studies should not only focus on the role of personality traits but also on people's appraisal of the situation as well as their emotional states, such as people's empathy with the person who treated them unfairly. Empathy with the transgressor makes people more likely to forgive after an apology (McCullough, Worthington, & Rachal, 1997) and perspective taking can be linked to lower rates of punishment in economic games (Will, Crone, & Grolzlu, 2015). Maybe these factors also affect the degree of people's transformative justice motives and the way people express them.

Similarly, there are at least two different ways in which people's implicit theories on moral character (whether it is malleable or not, see Schumann & Dweck, 2014) could affect transformative justice dynamics. On the one hand, entity theorists (i.e. individuals who consider moral character to be stable and fixed) and incremental theorists (i.e. individuals who consider moral character to be malleable) might be equally motivated to effect a change in the transgressor but express this motivation in different ways. It is possible that entity theorists are more likely than incremental theorists to use punishment as a tool to effect a change in the transgressor because entity theorists are more likely to attribute transgressions to internal, stable character traits that increase the likelihood of future transgressions (Tam, Shu, Ng, & Tong, 2013) whereas incremental theorists attribute transgressions to less stable character traits that might easily be changed by using non-punitive means. On the other hand, entity and incremental theorists might differ in their degree of transformative justice motives. Entity theorists react to transgressions with more negative affect than incremental theorists (Miller, Burgoon, & Hall, 2007) and might care less about change in transgressors than incremental theorists who believe that change is possible.

Interestingly, even the transgressor's implicit theory on moral character might matter, as empirical findings suggest that transgressors who see their own moral character as malleable (i.e. incremental theorists) are more likely to accept responsibility for what they did (Schumann & Dweck, 2014) and are presumably more likely to change. Thus, there could be even more potential for reconciliation if punishers made transgressors believe that their moral character is malleable.

Transformation as redemption? The role of spontaneous change
 Future studies could also focus on the role of transgressor change and its effects on people's justice-related satisfaction prior to the imposition of punishment to study when exactly punishment is needed for change to be satisfying. Findings from Studies 4 suggest that satisfaction does not differ substantially between people who decide to punish and people who purposefully decided to refrain from punishment. Punitive and merciful participants in these studies were equally satisfied after transgressor change. Findings from Study 1a, however, suggest that people who did not have to punish were less satisfied than punishers who received feedback about change. It remains to be examined if transformative motives can be addressed by spontaneous, yet genuine, change in the transgressor or only by change that results from the deliberate imposition of punishment (or the deliberate refraining from punishment) that has evoked a sense of guilt in the transgressor. Understanding when punishment or the possibility thereof is necessary to make change satisfying will clarify the link between transformative justice motives and notions of redemption (Garvey, 1999).

What are the factors that mediate the effects of transgressor change on justice-related satisfaction? In order to better understand people's transformative justice motives, it would be interesting to examine when participants' justice-related satisfaction after transgressor change is affected by the visibility of punishment and when not. The Needs-Based Model of Reconciliation suggests that victims aim for a restored feeling of empowerment (Shnabel & Nadler, 2008). Similarly, Justice Restoration Theory highlights the role of regained sense of status/power and restoration of values (Okimoto & Wenzel, 2008; Wenzel et al., 2008). If regaining

empowerment or restoring values are the reasons why transformational transgressors' attitudes are satisfying for punishers, one would expect to find the effect of transgressor change on justified satisfaction only if punishment is visible to the transgressor like in the present studies or to third parties. If transformative motives are not related to these concerns, however, visibility should not matter. Recent research findings suggest that people punish even if they know that the transgressor is unaware of the punishment (Crockett et al., 2014; Nadelhoffer, Heshmati, Kaplan, & Nichols, 2013). Thus, it might be that by punishing offenders in a 'hidden' way, punishers nonetheless hope to influence the offender's behavior or attitude indirectly, even without the transgressor's awareness of these intentions.

Conclusion

Findings from the present dissertation demonstrate that people have transformational justice motives after being treated unfairly that aim at a genuine change in the transgressor's attitude. Both punishers and nonpunishers hoped for a change in the transgressor, especially if they shared a social identity with the transgressor.

Current attempts that try to characterize taxonomies of punishment motives as focusing either on punitive or constructive means and as aiming at either consequentialist or deontological goals are not able to integrate the findings of the present set of studies (see, e.g., Wenzel & Okimoto, in press, Table 1, for an overview). Punitive means are usually linked to consequentialist goals like deterrence and incapacitation and deontological goals like retribution, whereas constructive means are usually linked to the consequentialist goal of rehabilitation or the deontological goal of restoration. The present findings suggest that people's behavior is not in line with such a

taxonomy of punishment philosophies, indeed non-punishers seem to aim at constructive rehabilitation. But sometimes people use punitive means to achieve constructive goals. They punish in order to effect a change in the transgressor. If such a change does not occur, people do not have a sense of restored justice.

Moreover, the present studies suggest that people often do not react with destructive, punitive means to achieve justice. Psychological theories on punishment need to be rooted within a broader framework of justice restoration and should not only look at either consequential or retributive punishment motives. The transformative justice framework provided in this dissertation would be a good place to start.

Taken together, retributive justice does not seem to be that retributive at all. It has transformative components that aim at reconciliation. These findings have implications for philosophical and psychological justice theories, as well as for legal policies. And lastly, they lead to many new interesting research questions that will hopefully inspire innovative research programs in the future.

References

- Aharoni, E., & Fridlund, A. J. (2012). Punishment without reason: Isolating retribution in lay punishment of criminal offenders. *Psychology, Public Policy, and Law*, 18(4), 599-625. doi: 10.1037/a0025821
- Ames, D. L., & Fiske, S. T. (2013). Intentional harms are worse when they're not. *Psychological Science*, 24(9), 1755-1762. doi: 10.1177/0956797613480507
- Ames, D. L., & Fiske, S. T. (2015). Perceived intent motivates people to magnify observed harm. *Proceedings of the National Academy of Sciences*, 112(12), 3599-3605. doi: 10.1073/pnas.1501592112
- Back, M. D., Kufner, A. C., Dufner, M., Gerlach, T. M., Rauthmann, J. F., & Denissen, J. J. (2013). Narcissistic admiration and rivalry: disentangling the bright and dark sides of narcissism. *Journal of Personality and Social Psychology*, 105(5), 1013-1037. doi: 10.1037/a0034431
- Baumeister, R. F., Stillwell, A. M., & Heatherton, T. F. (1994). Guilt: An interpersonal approach. *Psychological Bulletin*, 115(2), 243-267. doi: 10.1037/0033-2909.115.2.243
- Bazemore, G. (1998). Restorative justice and earned redemption: Communities, victims, and offender reintegration. *American Behavioral Scientist*, 41(6), 768-813.
- Bentham, J. (1830/2008). The rationale of punishment. In R. Smith (Ed.), *The making of the modern world*. Farmington Hills, MI: Thomson Gale.
- Bornstein, B. H., Rung, L. M., & Miller, M. K. (2002). The effects of defendant remorse on mock juror decisions in a malpractice case. *Behavioral Sciences & the Law*, 20(4), 393-409. doi: 10.1002/bsl.496

- Braithwaite, J. (1998). Restorative justice. In M. H. Tonry (Ed.) *The handbook of crime and punishment* (pp. 323-344). New York, NY US: Oxford University Press.
- Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high quality, data? *Perspectives on Psychological Science*, *6*(1), 3-5. doi: 10.1177/1745691610393980
- Bushman, B. J., & Baumeister, R. F. (1998). Threatened egotism, narcissism, self esteem, and direct and displaced aggression: Does *love* or selfhate lead to violence? *Journal of Personality and Social Psychology*, *75*, 219-229. doi: 10.1037/0022-3514.75.1.219
- Bushman, B. J., Baumeister, R. F., & Phillips, C. M. (2001). Do people aggress to improve their mood? Catharsis beliefs, affect regulation opportunity, and aggressive responding. *Journal of Personality and Social Psychology*, *81*, 117-32. doi: 10.1037/0022-3514.81.1.17
- Carlsmith, K. M. (2006). The roles of retribution and utility in determining punishment. *Journal of Experimental Social Psychology*, *42*, 437-451. doi: 10.1016/j.jesp.2005.06.007
- Carlsmith, K. M. (2008). On justifying punishment: The discrepancy between words and actions. *Social Justice Research*, *21*, 119-137. doi: 10.1007/s11211-008-0068-x
- Carlsmith, K. M., & Darley, J. M. (2008). Psychological aspects of retributive justice. In M. P. Zanna (Ed.), *Advances in Experimental Social Psychology* (Vol. 40, pp. 193-236). San Diego, CA: Elsevier.

- 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
- Carlsmith, K. M., Wilson, T. D., & Gilbert, D. T. (2008). The paradoxical consequences of revenge. *Journal of Personality and Social Psychology*, 95(6), 1316-1324. doi: 10.1037/a0012165
- 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
- Cullen, F. T., Pealer, A., Fisher, B. S., Applegate, B. K., & Santana, S. A. (2002). Public support for rehabilitation in America: Change or consistency. In J. V. Roberts & M. Hough (Eds.), *Changing attitudes to punishment* (pp. 128-147): Willian Publishing.
- Daly, K., & Proietti-Scifoni, G. (2011). Reparation and Restoration. In M. Tonry (Ed.), *The Oxford Handbook of Crime and Criminal Justice*. Oxford: Oxford University Press.
- Darby, B. W., & Schlenker, B. R. (1989). Children's reactions to transgressions: Effects of the actor's apology, reputation and remorse. *British Journal of Social Psychology*, 20(4), 353-364. doi: 10.1111/j.2044-8309.1989.tb00879.x
- Darley, J. M. (2002). Just punishments: Research on retributitional justice. In M. Ross & D. T. Miller (Eds.), *The justice motive in everyday life* (pp. 314-333). New York, NY US: Cambridge University Press.

- Darley, J. M. (2009). Morality in the law: The psychological foundations of citizens' desires to punish transgressions. *Annual Review of Law and Social Science*, 15, 1-23. doi: 10.1146/annurev.lawsocsci.4.110707.172335
- Darley, J. M., Carlsmith, K. M., & Robinson, P. H. (2000). Incapacitation and just deserts as motives for punishment. *Law and Human Behavior*, 24, 659-683. doi: 10.1023/A:1005552203727
- de Quervain, D. J. F., Fischbacher, U., Treyer, V., Schellhammer, M., Schnyder, U., Buck, A., & Fehr, E. (2004). The neural basis of altruistic punishment. *Science*, 305(5688), 1254-1258. doi: 10.1126/science.1100735
- Deigh, J. (2014). Punishment and proportionality. *Criminal Justice Ethics*, 33(3), 185-199. doi: 10.1080/0731129X.2014.988429
- Duff, R. A. (2001). *Punishment, Communication, and Community*. New York: Oxford University Press.
- Durkheim, É. (1893/1964). *The division of labor in society*. Glencoe: Free Press.
- Eaton, J., Ward Struthers, C., & Santelli, A. G. (2006). Dispositional and state forgiveness: The role of self-esteem, need for structure, and narcissism. *Personality and Individual Differences*, 41, 371-380. doi: 10.1016/j.paid.2006.02.005
- Eisenberg, T., Garvey, S. P., & Wells, M. T. (1998). But was he sorry? The role of remorse in capital sentencing. *Cornell Law Review*, 83(8).
- Ewing, A. C. (1927). Punishment as a moral agency: An attempt to reconcile the retributive and the utilitarian view. *Mind*, 36(143), 292-305. doi: 10.2307/2249539

- Ewing, A. C. (1929). *The morality of punishment*. London: Kegan Paul, Trench, Trubner & Co.
- Feather, N. T. (1999). *Values, achievement and justice: Studies in the psychology of deservingness*. New York: Kluwer Academic/Plenum Publishers.
- Fehr, E., & Gächter, S. (2002). Altruistic punishment in humans. *Nature*, 415(6868), 137-140. doi: 10.1038/415137a
- Feinberg, J. (1965). The expressive function of punishment. *The Monist*, 49(3), 397-423.
- Fisher, M. L., & Exline, J. J. (2006). Self-forgiveness versus excusing: The roles of remorse, effort, and acceptance of responsibility. *Self and Identity*, 5(2), 127-146. doi: 10.1080/15298860600586123
- Funk, F., McGeer, V., & Gollwitzer, M. (2014). Get the message: Punishment is satisfying if the transgressor responds to its communicative intent. *Personality and Social Psychology Bulletin*, 40(9), 986-997. doi: 10.1177/0146167214533130
- Garvey, S. P. (1998). Can shaming punishments educate? *The University of Chicago Law Review*, 65(3), 733-794.
- Garvey, S. P. (1999). Punishment as atonement. *UCLA Law Review*, 46(1), 180-185.
- Gerlach, T. M., Allemand, M., Agroskin, D., & Denissen, J. J. A. (2012). Justice sensitivity and forgiveness in close interpersonal relationships: the mediating role of mistustful, legitimizing, and pro-relationship cognitions. *Journal of Personality*, 80(5), 1373-1413. doi: 10.1111/j.1467-494.2012.00762.x
- Goffman, E. (1971). *Relations in public; microstudies of the public order*. New York: Basic Books.

- Gold, G. J., & Weiner, B. (2000). Remorse, confession, group identity, and expectancies about repeating a transgression. *Basic and Applied Social Psychology*, 22(4), 291-300. doi: 10.1207/15324830051035992
- Gollwitzer, M., & Bücklein, K. (2007). Are 'we' more punitive than 'me'? Self-construal styles, justice-related attitudes, and punitive judgments. *Social Justice Research*, 20(4), 457-478. doi: 10.1007/s11211-007-0051-y
- Gollwitzer, M., & Bushman, B. J. (2012). Do victims of injustice punish to improve their mood? *Social Psychological and Personality Science*, 5(5), 572-580. doi: 10.1177/1948550611430552
- Gollwitzer, M., & Denzler, M. (2009). What makes revenge sweet: Seeing the offender suffer or delivering a message? *Journal of Experimental Social Psychology*, 45(4), 840-844. doi: 10.1016/j.jesp.2009.03.001
- Gollwitzer, M., & Keller, L. (2010). What you did only matters if you are one of us: Offenders' group membership moderates the effect of criminal history on punishment severity. *Social Psychology*, 41(1), 20-26. doi: 10.1027/1864-9335/a000004
- Gollwitzer, M., Meder, M., & Schmitt, M. (2011). What gives victims satisfaction when they seek revenge? *European Journal of Social Psychology*, 41(3), 364-374. doi: 10.1002/ejsp.782
- Gollwitzer, M., & Rothmund, T. (2009). When the need to trust results in unethical behavior: The Sensitivity to Mean Intentions (SeMI) model. In D. D. Cremer (Ed.), *Psychological perspectives on unethical behavior and decision making* (pp. 135-152). Charlotte, NC: Information Age.

- Gollwitzer, M., Rothmund, T., Alt, B. & Jekel, M. (2012). Victim sensitivity and the accuracy of social judgments. *Personality and Social Psychology Bulletin*, 38(8), 975-984. doi: 10.1177/0146167212440887
- Gollwitzer, M., Rothmund, T., & Süssenbach, P. (2013). The Sensitivity to Mean Intentions (SeMI) Model: Basic assumptions, recent findings, and potential avenues for future research. *Social and Personality Psychology Compass*, 7(3), 415-426. doi: 10.1111/spc3.12041
- Goodwin, G. P., Piazza, J., & Rozin, P. (2014). Moral character predominates in perception and evaluation. *Journal of Personality and Social Psychology*, 106(1), 148-168. doi: 10.1037/a0034726
- Gorringer, T. (1996). *God's just vengeance: Crime, violence, and the rhetoric of salvation*. Cambridge: Cambridge University Press.
- Hampton, J. (1984). The moral education theory of punishment. *Philosophy & Public Affairs*, 13(3), 208-238. doi: 10.2307/2265412
- Hampton, J. (1992). Correction harms versus righting wrongs: The goal of retribution. *UCLA Law Review*, 39(1), 659-1702.
- Heider, F. (1958). *The psychology of interpersonal relations*. New York: Wiley & Sons.
- Henrich, J., McElreath, R., Barr, A., Ensminger, J., Barrett, C., Botyanatz, A., . . . Ziker, J. (2006). Costly punishment across human societies. *Science*, 312(5781), 1767-1770. doi: 10.1126/science.1125156
- Hofmann, W., Wisneski, D. C., Brandt, M. J., & Skitka, L. J. (2014). Morality in everyday life. *Science*, 345(6202), 1340-1343. doi: 10.1126/science.1251560

- Hogg, M. A., & Abrams, D. (1988). *Social identifications: A social psychology of intergroup relations and group processes*. London: Routledge.
- Hosser, D., Windzio, M., & Greve, W. (2008). Guilt and shame as predictors of recidivism: A longitudinal study with young prisoners. *Criminal Justice and Behavior*, 35(1), 138-152. doi: 10.1177/0093854807309224
- Kant, I. (1797/1968). *Die Metaphysik der Sitten* [The metaphysics of conventions]. *Kants Werke* (Akademie Textausgabe, Vol. 5). Berlin, Germany: Walter de Gruyter.
- Keller, L., Oswald, M., Stucki, I., & Gollwitzer, M. (2010). A closer look at an eye for an eye: Laypersons' punishment decisions are primarily driven by retributive motives. *Social Justice Research*, 23(3), 99-116. doi: 10.1007/s11211-010-0113-4
- Krasnow, M. M., Cosmides, L., Pedersen, E. J., & Tooby, J. (2012). At we punishment and reputation for. *PLoS ONE*, 7(9). doi: 10.1371/journal.pone.0045662
- Lachman, M. E., & Weaver, S. L. (1998). The sense of control as a moderator of social class differences in health and well-being. *Journal of Personality and Social Psychology*, 74(3), 763-773. doi: 10.1037/0022-3514.74.3.763
- Langner, O., Dotsch, R., Bijlstra, G., Wigboldus, D. H. J., Hawk, S. T., & van Knippenberg, A. (2010). Presentation and validation of the Radboud Faces Database. *Cognition & Emotion*, 24(8), 1377-1388. doi: 10.1080/02699930903485076
- Lazare, A. (2004). *On apology*. New York: Oxford University Press.

- Leach, C. W., van Zomeren, M., Zebel, S., Vliek, M. L., Pennekamp, S. F., Doosje, B., . . . Spears, R. (2008). Group-level self-definition and self-investment: A hierarchical (multicomponent) model of group identification. *Journal of Personality and Social Psychology*, *95*, 144-165. doi: 10.1037/0022-3514.95.1.144
- Lerner, M. J., & Clayton, S. (2011). *Justice and self-interest: Two fundamental motives*. Cambridge, UK: Cambridge University Press.
- Leyens, J.P., Rodriguez-Perez, A., Rodriguez-Torres, R., Gaunt, R., Paladino, M., Vaes, J., & Demoulin, S. (2001). Psychological essentialism and the differential attribution of uniquely human emotions to ingroups and outgroups. *European Journal of Social Psychology*, *31*, 395-411. doi: 10.1002/ejsp.50
- Lieberman, D., & Linke, L. (2007). The effect of social category on third party punishment. *Evolutionary Psychology*, *5*, 289-305.
- Malle, B. F., Monroe, A. E., & Gaglielmo, S. (2014). Paths to blame and paths to convergence. *Psychological Inquiry*, *25*, 251-260. doi: 10.1080/1047840X.2014.913379
- Marques, J. M., Abrams, D., & Serodio, R. G. (2001). Being better by being right: Subjective group dynamics and derogation of in-group deviants when generic norms are undermined. *Journal of Personality and Social Psychology*, *81*, 436-447. doi: 10.1037/0022-3514.81.3.436
- Marques, J. M., Yzerbyt, V. Y., & Leyens, J. (1988). The 'Black Sheep Effect': Extremity of judgments towards ingroup members as a function of group

- identification. *European Journal of Social Psychology* (18), 1-16. doi: 10.1002/ejsp.2420180102
- Martin, J. W., & Cushman, F. A. (in press). The adaptive logic of moral luck. In J. Sytsma & W. Buckwalter (Eds), *The Blackwell Companion to Experimental Philosophy*
- McAdams, D. P. (2013). *The redemptive self: Stories Americans live by* (Revised and Expanded Edition). New York, NY: Oxford University Press.
- McCullough, M. E. (2008). *Beyond revenge: The evolution of the forgiveness instinct*. San Francisco: Jossey-Bass.
- McCullough, M. E., Worthington, E. L., & Rachal, K. C. (1997). Interpersonal forgiving in close relationships. *Journal of Personality and Social Psychology* (73), 321-336. doi: 10.1037/0022-3514.73.2.321
- McGeer, V. (2013). Civilizing blame. In J. D. Coates & N. A. Tognazzini (Eds), *Blame: Its nature and norms* (pp. 162-188). Oxford: Oxford University Press.
- 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*. doi: 10.1177/1948550614527115
- Miller, C. H., Burgoon, J. K., & Hall, J. R. (2007). The effects of implicit theories of moral character on affective reactions to moral transgressions. *Social Cognition*, 25(6), 819-832. doi: 10.1521/soco.2007.25.6.819
- Miller, D. T. (2001). Disrespect and the experience of injustice. *Annual Review of Psychology*, 52, 527-553. doi: 10.1146/annurev.psych.52.1.527
- Moberly, W. (1968). *The ethics of punishment*. London: Faber & Faber.

- Morris, H. (1981). A paternalistic theory of punishment. *American Philosophical Quarterly*, 18(4), 263-271. doi: 10.2307/20013924
- Nadelhoffer, T., Heshmati, S., Kaplan, D., & Nichols, S. (2013). Folk retributivism and the communication confound. *Economics & Philosophy*, 29(2), 235-261. doi: 10.1017/S0266267113000217
- Neuberg, S. L., & Newsom, J. T. (1993). Personal need for structure: Individual differences in the desire for simpler structure. *Journal of Personality and Social Psychology*, 65(1), 113-131. doi: 10.1037/0022-3514.65.1.113
- Nozick, R. (1981). *Philosophical explanations*. Cambridge, MA: Harvard University Press.
- Nugier, A., Chekroun, P., Pierre, K., & Niedenthal, P. M. (2009). Group membership influences social control of perpetrators of uncivil behavior. *European Journal of Social Psychology*, 39(9), 1126-1134. doi: 10.1002/ejsp.602
- Ohtsubo, Y., & Watanabe, E. (2009). Do sincere apologies need to be costly? Test of a costly signaling model of apology. *Evolution and Human Behavior*, 30(1), 114-123. doi: 10.1016/j.evolhumbehav.2008.09.004
- Okimoto, T. G., & Wenzel, M. (2008). The symbolic meaning of transgressions: Towards a unifying framework of justice restoration. In K. A. Heggie & J. Clay-Warner (Eds.), *Advances in Group Processes* (Vol. 25, pp. 291-326). Oxford, UK: Elsevier.
- Okimoto, T. G., & Wenzel, M. (2009). Punishment as restoration of group and offender values following a transgression: value consensus through symbolic holding

- and offender reform *European Journal of Social Psychology*, (39), 346-367.
doi: 10.1002/ejsp.537
- Okimoto, T. G., & Wenzel, M. (2014). Bridging diverging perspectives and repairing damaged relationships in the aftermath of workplace transgressions. *Business Ethics Quarterly*, 24(3), 443-473. doi: 10.5840/beq201471515
- Okimoto, T. G., Wenzel, M., & Feather, N. T. (2012). Retribution and restoration as general orientations towards justice. *European Journal of Personality*, 26, 255-275. doi: 10.1002/per.831
- Oppenheimer, D. M., Meyvis, T., & Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology*, 45, 867-872. doi: 10.1016/j.jesp.2009.03.009
- Orth, U. (2003). Punishment goals of crime victims. *Law and Human Behavior*, 27, 173-186. doi: 10.1023/A:1022547213760
- Orth, U. (2004). Does perpetrator punishment satisfy victims' feelings of revenge? *Aggressive Behavior*, 30, 62-70. doi: 10.1002/ab.20003
- Oswald, M.E., Hupfeld, J., Klug, S. C., & Gabriel, U. (2002). Lay-perspectives on criminal deviance, goals of punishment, and punitiveness. *Social Justice Research*, 15(2), 85-98. doi: 10.1023/A:1019928721720
- Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making*, 5, 411-419.
- Paysan, P., Knothe, R., Amberg, B., Romdhani, S., & Vetter, T. (2009). A 3D face model for pose and illumination invariant face recognition. *Proceedings of the*

- 6th IEEE International Conference on Advanced Video and Signal based Surveillance (AVSS) for Security, Safety and Monitoring in Smart Environments
- Petersen, M. B., Sell, A., Tooby, J., & Cosmides, L. (2010). Evolutionary psychology and criminal justice: A recalibrational theory of punishment and reconciliation. In H. Hogh-Oleson (Ed.) *From human morality & sociality: Evolutionary and comparative perspectives* (pp. 721-131). New York: Palgrave MacMillan.
- Pipes, R. B., & Alessi, M. (1999). Remorse and a previously punished offense in assignment of punishment and estimated likelihood of a repeated offense. *Psychological Reports*, 85, 246-248. doi: 10.2466/pr0.1999.85.1.246
- Proeve, M., & Tudor, S. (2010). *Remorse Psychological and jurisprudential perspectives*. Burlington, VT: Ashgate.
- Robinson, D. T., Smith-Lovin, L., & Tsoudis, O. (1994). Heinous crime or unfortunate accident? The effects of remorse on responses to mock criminal confessions. *Social Forces*, 73(1), 175-190. doi: 10.2307/2579922
- Rothman, D. J. (1971). *The discovery of the asylum*. Boston: Little Brown.
- Rumsey, M. C. (1976). Effects of defendant background and remorse on sentencing judgments. *Journal of Applied Social Psychology*, 16, 64-68. doi: 10.1111/j.1559-1816.1976.tb01312.x
- Scher, S., & Darley, J. (1997). How effective are the things people say to apologize? Effects of the realization of the apology speech. *Journal of Psycholinguistic Research*, 26(1), 127-140. doi: 10.1023/A:1025068306386

- Schmitt, M., Gollwitzer, M., Maes, J., & Arbach, D. (2005). Justice Sensitivity: Assessment and location in the personality space. *European Journal of Psychological Assessment*, 21, 202-211. doi: 10.1027/10155759.21.3.202
- Schmitt, M., Neumann, R., & Montada, L. (1995). Dispositional sensitivity to befallen injustice. *Social Justice Research*, 14, 385-407. doi: 10.1007/BF02334713
- Schumann, K., & Dweck, C. S. (2014). Who accepts responsibility for their transgressions? *Personality and Social Psychology Bulletin*, 40, 1598-1610. doi: 10.1177/0146167214552789
- Shiffman, S., Stone, A. A., & Hufford, M. R. (2008). Ecological momentary assessment. *Annu Rev Clin Psychol*, 4, 1-32.
- Shinada, M., Yamagishi, T., & Ohmura, Y. (2004). False friends are worse than bitter enemies: 'Altruistic' punishment of ingroup members. *Evolution and Human Behavior*, 25, 379-393. doi: 10.1016/j.evolhumbehav.2004.08.001
- Shnabel, N., & Nadler, A. (2008). A needs-based model of reconciliation: Satisfying the differential emotional needs of victim and perpetrator as a key to promoting reconciliation. *Journal of Personality and Social Psychology*, 94, 116-132. doi: 10.1037/0022-3514.94.1.116
- Slovenko, R. (2006). Remorse. *The Journal of Psychiatry & Law*, 34, 429-432.
- Smith, A. (1759/1869). *A theory of moral sentiments*. London: Ball & Daldy.
- Tajfel, H., & Turner, J. (2001). An integrative theory of intergroup conflict. In M. A. Hogg & D. Abrams (Eds.), *Intergroup relations: Essential readings* (pp. 94-109). New York, NY US: Psychology Press.

- Tam, K.-P., Shu, T.-M., Ng, H. K.-S., & Tong, Y.-Y. (2013). Belief about immutability of moral character and punitiveness toward criminal offenders. *Journal of Applied Social Psychology*, 43, 603-611. doi: 10.1111/j.1559-1816.2013.01041.x
- ten Brinke, L., MacDonald, S., Porter, S., & O'Connor, B. (2012). Crocodile tears: Facial, verbal and body language behaviours associated with genuine and fabricated remorse. *Law and Human Behavior*, 36, 51-59. doi: 10.1037/h0093950
- Tsoudis, O., & Smith-Lovin, L. (1998). How bad was it? The effects of victim and perpetrator emotion on responses to criminal court vignettes. *Social Forces*, 77(2), 695-722. doi: 10.2307/3005544
- Uhlmann, E. L., Pizarro, D. A., & Diermeier, D. (2015). A person-centered approach to moral judgment. *Perspectives on Psychological Science*, 11, 72-81. doi: 10.1177/1745691614556679
- van Prooijen, J.W. (2006). Retributive reactions to suspected offenders: The importance of social categorizations and guilt probability. *Personality and Social Psychology Bulletin*, 32(6), 715-726. doi: 10.1177/0146167205284964
- Vidmar, N. (2001). Retribution and revenge. In V. L. Hamilton (Ed.), *Handbook of justice research in law* (pp. 31-63). New Haven, CT: Yale University Press.
- Vidmar, N. (2002). Retributive justice as social context. In M. Ross & D. T. Miller (Eds.), *The justice motive in everyday life* (pp. 291-313). New York, NY US: Cambridge University Press.

- Vidmar, N., & Miller, D. T. (1980). Social psychological processes underlying attitudes toward legal punishment. *Law & Society Review*, 14, 401-439.
- Walker, M., & Vetter, T. (2009). Portraits made to measure: Manipulating social judgments about individuals with a statistical face model. *Journal of Vision*, 9(11), 1-13. doi: 10.1167/9.11.12
- Wallace, H. M., Exline, J. J., & Baumeister, R. F. (2008). Interpersonal consequences of forgiveness: Does forgiveness deter or encourage repeat offenses? *Journal of Experimental Social Psychology*, 44, 453-460. doi: 10.1016/j.jesp.2007.02.012
- Wenzel, M., & Okimoto, T. G. (in press). Retributive justice. In C. Sabbagh & M. Schmitt (Eds.) *Handbook of social justice theory and research*. Springer Press.
- Wenzel, M., Okimoto, T. G., Feather, N. T., & Platow, M. J. (2008). Retributive and restorative justice. *Law and Human Behavior*, 32(5), 375-389. doi: 10.1007/s10979-007-9116-6
- Wenzel, M., & Thielmann, I. (2006). Why we punish in the name of justice: Just desert versus value restoration and the role of social identity. *Social Justice Research*, 19(4), 450-470. doi: 10.1007/s11211-006-0028-2
- Will, G.-J., Crone, E. A., & Grožl, B. (2015). Acting on social exclusion: neural correlates of punishment and forgiveness of exclusion. *Social Cognitive and Affective Neuroscience*, 10, 209-218. doi: 10.1093/scan/nsu045
- Wohl, M. J. A., Horsey, M. J., & Bennett, S. H. (2012). Why group apologies succeed and fail: Intergroup forgiveness and the role of primary and secondary emotions.

Journal of Personality and Social Psychology, (10), 2306-322. doi:
10.1037/a0024838

Woodyatt, L., & Wenzel, M(2013). The psychological immune response in the face of transgressions: Pseudo self-forgiveness and threat to belonging. Journal of Experimental Social Psychology, (49), 951-958. doi:
10.1016/j.jesp.2013.05.016

Xiao, E., & Houser, D. (2005). Emotion expression in human punishment behavior. Proceedings Of The National Academy Of Sciences Of The United States Of America, 102(20), 7398-7401.

Zehr, H. (1990) Changing lenses: A new focus for crime and justice. Scottsdale, PA: Herald Press.

Appendix

Feedback conditions Study 2

No change/ no suffering

No change/ suffering

Mere behavior change

Change/ no suffering

Change/ suffering

All pictures were generated online using Cambridge English Online's cartoon maker:
http://cambridgeenglishonline.com/Cartoon_Maker/

Individual difference measures used in Studies 2 and 4

Victim sensitivity (Schmitt et al., 2005)

It bothers me when others receive something that ought to be mine; It makes me angry when others receive an award which I have earned; I can't easily bear it when others profit unilaterally from me; I can't forget for a long time when I have to fix others' carelessness; It gets me down when I get fewer opportunities than others to develop my skills; It makes me angry when others are undeservingly better off than me; It worries me when I have to work hard for things that come easily to others; I ruminate for a long time when other people are being treated better than me; It burdens me to be criticized for things that are being overlooked with others; It makes me angry when I am treated worse than others (1 = do not agree at all 7 = completely agree)

Sense of control (taken from Lachman & Weaver, 1998)

Whether or not I am able to get what I want is in my own hands; What happens to me in the future mostly depends on me; What happens in my life is often beyond my control (-); I sometimes feel I am being pushed around in my life (1 = do not agree at all; 7 = completely agree)

Personal need for structure (Neuberg & Newsom, 1993)

It upsets me to go into a situation without knowing what I can expect from it; I'm not bothered by things that interrupt my daily routine; I enjoy having a clear and structured mode of life; I like to have a place for everything and everything in its place; I enjoy being spontaneous (-); I find that a wellordered life with regular hours makes my life tedious (-); I don't like situations that are uncertain; I hate to change my plans at the last minute; I hate to be with people who are unpredictable; I find that a consistent

routine enables me to enjoy life more; I enjoy the exhilaration of being in unpredictable situations(-); I become uncomfortable when the rules in a situation are not clear (1 = strongly disagree, 7 = strongly agree)

Narcissistic Admiration and Rivalry Questionnaire (Back et al., 2013), only used in Study 2

Subscale Admiration. I am great; I will someday be famous; I deserve to be seen as a great personality; I show others how special I am; I enjoy myself very much; Being a very special person gives me a lot of strength; Most of the time I am able to draw people's attention to myself in conversations; I manage to be the center of attention with my outstanding contributions; Mostly, I am very adept at dealing with other people.

Subscale Rivalry. Most people won't achieve anything; Other people are worth nothing; Most people are somehow losers; I secretly take pleasure in the failure of my rivals; I want my rivals to fail; I enjoy it when another person is inferior to me; I react annoyed if another person steals the show from me; I often get annoyed when I am criticized; I can barely stand it if another person is at the center of events (1 = do not agree at all, 7 = agree completely)

Ingroup identification scale (Leach et al., 2008) adapted to Princeton context, only used in Study 4

I feel committed to Princeton undergraduate students; It is pleasant to be a Princeton undergraduate student; Being a Princeton undergraduate student is an important part of how I see myself; I have a lot in common with the average Princeton

undergraduate student; Princeton undergraduate students have a lot in common with each other (1 = strongly disagree, 7 = strongly agree).

Self-construal scale (Gollwitzer & Buecklein, 2007) „ only used in Study 4

Subscale Interdependent/ context dependence My life is seldom influenced by other people (-), When making decisions, I let other people influence me; In general, other people's behavior influences my own behavior; Generally, my life is, in many aspects, influenced by other people; My friends' opinions are in general influential on my own beliefs (-); My well-being frequently depends on people that are important to me; My self-image largely depends on how close friends perceive me.

Subscale Interdependent/ relatedness I spend a lot of time on my friendships; Even when I'm currently busy, I make time for my friends; Friendships make up a great deal of my personality; It is always very important to me that the people who are close to me are feeling good; I am often happy for other people when they are successful.

Subscale Independence I like being unique and in many respects different from others; Usually, I set a high value on things that differentiate me from other people; Individualism always plays an important role in my life; I possess many characteristics that others do not possess (1 = not at all true, 7 = totally true).

Items within each scale were presented in randomized order.

(-) indicates that the item is recoded.