British Journal of Social Psychology (2019), 58, 869–893

The British Psychological Society

www.wileyonlinelibrary.com

Justice reactions to deviant ingroup members: Ingroup identity threat motivates utilitarian punishments

Kyriaki Fousiani[1]*, Vincent Yzerbyt[2], Nour-Sami Kteily[3] and Stéphanie Demoulin[2]

[1]Faculty of Behavioural and Movement Sciences, Vrije Universiteit Amsterdam, The Netherlands

[2]Faculté de psychologie et des sciences de l'éducation, Université catholique de Louvain-La-Neuve, Belgium

[3]Kellogg School of Management, Northwestern University, Evanston, Illinois, USA

To maintain a positive overall view of their group, people judge likeable ingroup members more favourably and deviant ingroup members more harshly than comparable outgroup members. Research suggests that such derogation of deviant ingroup members aims to restore the image of the group by symbolically excluding so-called 'black sheeps'. We hypothesized that information about a harm-doer's group membership influences observers' justice-seeking reactions. Motives for punishment vary based on whether the goal is to punish past harm-doing (i.e., retributive motives), help harm-doers recognize the harm inflicted and reintegrate into society (i.e., restorative motives), or control harm-doer's future behaviour through incapacitating practices and exclusion from society (i.e., utilitarian motives). We hypothesized that immoral behaviours by ingroup rather than outgroup members jeopardize the group's reputation and therefore activate utilitarian (i.e., exclusion-oriented) motives for punishment. Study 1 (N = 187) confirmed that people displayed more utilitarian motives and less restorative motives when sanctioning an ingroup as opposed to an outgroup harm-doer. Study 2 (N = 122) manipulated typicality to the ingroup. Participants displayed stronger utilitarian (i.e., exclusion-oriented) punishment motives when the harm-doer was presented as a typical ingroup rather than an outgroup member. Study 3 (N = 292) replicated the findings of Studies 1 and 2 and further showed that people displayed stronger utilitarian punishments against an ingroup offender through the experience of increased identity threat. Contrary to our expectations, observers' ingroup identification did not moderate the effect of group membership or typicality to the ingroup on justice reactions. Yet, ingroup identification influenced both experienced identity threat (i.e., mediator) and utilitarian motives for punishment with high identifiers experiencing higher threat and displaying stronger utilitarian punishment motive. We discuss the results in terms of people's concern for the protection of their group identity.

**Correspondence should be addressed to Kyriaki Fousiani, Faculty of Behavioural and Movement Sciences, Vrije Universiteit Amsterdam, Van der Boechorststraat 7, 1081 BT, Amsterdam, The Netherlands (email: k.fousiani@vu.nl).*

DOI:10.1111/bjso.12312

Morality is highly valued (Haslam, 2006, 2015) and people wish to belong to groups that they and others see as being moral (Ellemers & Van den Bos, 2012; Leach, Ellemers, & Barreto, 2007). Indeed, morality is a major source of group pride (Branscombe, Ellemers, Spears, & Doosje, 1999; Leach *et al.*, 2007) and immoral doing is experienced as aversive (Branscombe, Doosje, & McGarty, 2002). For instance, immorality, when displayed by ingroup as compared to outgroup members, is associated with ingroup-directed hostility through collective shame (Piff, Martinez, & Keltner, 2012), psychological distress, and experience of threat to shared values (De Castella, Platow, Wenzel, Okimoto, & Feather, 2011; Okimoto & Wenzel, 2010; Rullo, Presaghi, & Livi, 2015; Sankaran, Sekerdej, & Von Hecker, 2017; Van der Toorn, Ellemers, & Doosje, 2015). Additionally, immorality is viewed as the greatest 'threat to the image' of one's group (Brambilla, Sacchi, Pagliaro, & Ellemers, 2013; Pagliaro, Brambilla, Sacchi, D'Angelo, & Ellemers, 2013).

Research on the 'black sheep effect' (BSE; Marques & Yzerbyt, 1988; Marques, Yzerbyt, & Leyens, 1988) shows that when a threat to one's group identity comes from the inside, that is, from an ingroup member, people respond by treating the ingroup transgressor very negatively. In fact, compared to outgroup transgressors, the derogation of ingroup transgressors can be much harsher. This happens because an immoral behaviour by ingroup members jeopardizes the reputation and the image of the ingroup (Brambilla *et al.*, 2013), inconsistent with the perceived duty of ingroup members to represent the group in the best possible light (Hornsey *et al.*, 2005) and increases ingroup identity threat (Van der Toorn *et al.*, 2015). Symbolic exclusion of deviant ingroup members thus aims to restore the harmed image of the group (Marques & Yzerbyt, 1988; Marques *et al.*, 1988) and seems a more efficient strategy compared to downplaying or denying the transgression (Van Leeuwen, van den Bosch, Castano, & Hopman, 2010).

Another important variable that should impact how group members are viewed is their ingroup typicality – that is their representativeness of the ingroup. Typical members, that is, member who have sufficiently integrated the core values and beliefs of their group and understand the importance of conforming to group norms (Abrams, Palmer, Rutland, Cameron, & Van de Vyver, 2014), come across as more representative, core members of the group. These members play a special role in validating the group's social identity, and their deviation from the group's standards is considered as more serious as it poses a more significant threat to the group's identity (Levine & Moreland, 2002). Deviations from the ingroup values and norms appear less severe or threatening when displayed by marginal/atypical members of the group. In a similar vein, Meeussen *et al.* (2012) found that deviance by typical ingroup members triggers increased ingroup identity threat, whereas deviance by less typical ingroup members generates lower levels of experienced ingroup identity threat. Pinto, Marques, Levine, and Abrams (2010) found that typical ingroup members (i.e., full ingroup members who have successfully completed socialization and undergone the role transition of acceptance) who deviate suffer more derogation compared to deviant marginal (i.e., members who once held the status of full members but lost this status because they failed to live up to the group's expectations) or new members (i.e., members who have provisional status in the group and must prove themselves to old-timers in order to make the role transition of acceptance). Moreover, derogation of typical ingroup members was associated with more punishing intention compared to derogation of deviant marginal ingroup members. This presumably happens because typical members of the ingroup are responsible for upholding the group's core beliefs and transmitting them to the new members (Levine & Moreland, 1994). In other words, typical members legitimize the group's values when they behave properly but jeopardize them when they deviate from the norms.

Interestingly, not all people view their group membership as equally significant for their self-perception and thus not everyone reacts similarly against ingroup deviants. People who are highly identified define themselves in terms of their group membership to a greater degree than those who identify less strongly. Several studies demonstrate that highly identified group members react especially positively towards ingroup members who help the group achieve a positive social identity. In contrast, they react particularly negatively towards ingroup members who reflect poorly on the group (Abrams, Marques, Bown, & Henson, 2000; Branscombe, Wann, Noel, & Coleman, 1993; Hutchison & Abrams, 2003). Accordingly, high identifiers perceive ingroup deviants as less typical of the ingroup, which seems to allow them, in turn, to safeguard the image of their group (Castano, Paladino, Coull, & Yzerbyt, 2002). As the group's reputation and image are less self-relevant for the low identified and have less of their self invested in the ingroup, they often react similarly towards ingroup members whether those reflect positively or negatively on the group (Hutchison & Abrams, 2003). Research thus suggests that it is important to consider the role of ingroup identification as a moderator on reactions to deviant ingroup behaviours.

Prior work has looked at the way people punish ingroup versus outgroup members (Braun & Gollwitzer, 2012; Brown, González, Zagefka, Manzi, & Čehajić, 2008; Gollwitzer & Keller, 2010; Gollwitzer & Van Prooijen, 2016; Van Prooijen & Lam, 2007). For instance, individuals tend to punish repeat ingroup offenders more severely than first-time ingroup offenders, whereas criminal history does not affect punitive reactions towards outgroup offenders (Gollwitzer & Keller, 2010). Similarly, Braun and Gollwitzer (2012) indicated that applying harsher punishments to ingroup rather than outgroup offenders often serves as a means by which ingroup members seek to protect their group's image. However, there is, to the best of our knowledge, no research examining what people's motives for punishing the ingroup versus outgroup transgressors are. Indeed, motives for punishment can vary from rehabilitative/restorative (Zehr, 1997) to proportional (Goldberg, Lerner, & Tetlock, 1999; Kant, 1797) or zero-tolerance (Darley, Carlsmith, & Robinson, 2000). In the present paper, we investigate the role of offenders' group membership (ingroup vs. outgroup) and perceived ingroup typicality (typical to the ingroup vs. atypical to the ingroup) on justice decisions, taking into account differential motives for distribution of justice (Carlsmith, 2008; Carlsmith, Darley, & Robinson, 2002; Darley *et al.*, 2000). In addition, we investigate the mediating role of ingroup identity threat in the relationship between offenders' group membership and ingroup typicality, and distribution of justice (i.e., punishment motives). Finally, we explore whether the aforementioned relationship varies as a function of the observer's identification with the ingroup.

Motives behind punishing reactions

Exploring how harsh or lenient punishments against transgressors can be does not give any information about the goals that punishers wish to achieve through their justice decisions. Punishment can serve a variety of goals, and it is important to gain a better understanding of the message people wish to send by selecting particular forms of punishment. Work has identified three distinct kinds of motives for punishment: utilitarian motives (Bentham, 1789), retributive or just deserts motives (Kant, 1797), and restorative or rehabilitation motives (de Beaumont & Tocqueville, 1833; Saleilles, 1898). Each of these motives serves a different goal and selecting one or another form of punishment has important consequences for the offender.

Utilitarian motives for punishment aim to reduce the likelihood of offences in the future and thus at the maximization of happiness and minimization of suffering among many other people (see Carlsmith & Darley, 2008; Nagin, 1998; Van Prooijen, 2018). The purpose of utilitarian punishments is to control the offenders' behaviour through deterrence of future crimes or incapacitation of a known liability to society (Carlsmith & Darley, 2008). Deterrence assumes that the offender is a rational person who has designed and deliberately committed an offence, and calculated the expected benefits in relation to the expected costs (e.g., likelihood and severity of punishment). Incapacitation assumes that the offender is unable to act rationally and needs to be restrained, into a legal quarantine (e.g., prison) so that the prevalence of crime is reduced. Utilitarian punishment implies a zero-tolerance punishment (Nagin, 1998) and traditionally includes sentences like long-time incarceration, capital punishment, deportation, or disbarment (see Carlsmith & Darley, 2008, pp. 200–201). An aspect well integrated into the concept of utilitarianism is *instrumental harm*, which favours the acceptance of instrumental use of people for the maximization of greater good (see also Kahane *et al.* (2018).

In contrast to utilitarian punishing practices, *retributive/just deserts* punishments rest on the moral philosophy of deontology according to which punishment must be proportionate to the harm inflicted. Retributive punishment's objective is not preventing future offences per se, but retaliating for perpetrators' past behaviour (Goldberg *et al.*, 1999; see also Van Prooijen, 2018). As put by Kant (1797): 'punishment can never be administered merely as a means for promoting another good, and should be pronounced over all criminals proportionate to the internal wickedness' (p. 397). People are more likely to trust retribution (also seen as deontology-based), which rejects harming one person for the benefit of many others, and distrust utilitarianism (also called 'consequentialist'). However, it should be noted that utilitarianism, except of instrumental harm, includes a positive aspect as well, that is, impartial beneficence (Kahane *et al.*, 2018), through which, one should expect punishments to be free of bias or favouritism towards one's social network or personal ties (see Everett, Faber, Savulescu, & Crockett, 2018).

Finally, and in addition to the more traditional punitive approaches, justice-seeking reactions directed at harm-doers can also involve motives of *restoration or rehabilitation* (de Beaumont & Tocqueville, 1833; Saleilles, 1898). From this perspective, the needs of both the victims and the offenders enter the picture. The core element of restoration is to encourage offenders to take responsibility for their actions and to 'repair' the harm inflicted. This can be achieved through several means that include, for instance, apology, community service, returning of stolen objects-money, and so on. Restorative justice focuses on the harm-doer as a person and not on the harm as an action that requires punishment. It emphasizes the need to help harm-doers recognize the harm they have caused, to have them apologize to the victim and repair the relationship between the harm-doer and the victim, and to alter the harm-doer's future behaviour by means of adequate treatment (Zehr, 1997).

Aims of the study and hypotheses

Three experiments examined how group-based characteristics of immoral doers affect the motives under which observers assign punishments. More specifically, we investigated the effect of offenders' group characteristics on perceivers' selection of proportionate, inclusion, or exclusion-oriented punishing practices.

First, according to the justice literature, regardless of any group-related characteristics of the offender, observers are expected to display a generalized tendency to assign more retributive and less utilitarian or restorative motives for punishing the harm-doer. Indeed, people have a stronger intuitive need to punish *past* transgressions than to punish in order to prevent future harm-doing (Wenzel & Thielmann, 2006; see also Carlsmith, 2008; Carlsmith *et al.*, 2002; Darley & Pittman, 2003). Moreover, they display an intuitive preference for proportionate rather than extreme sanctions. In other words, retribution is the type of punishment that aims at rebalancing feelings of justice and fairness. We thus expect that this type of punishment will be assigned to a higher extent than utilitarian or restorative motives for sanctioning (*Hypothesis 1*).

Second, based on the BSE theory (Marques & Yzerbyt, 1988; Marques *et al.*, 1988), we expect that individuals' motives for punishment should vary depending on the group membership of the offender. Because immorality enacted by ingroup members is considered as a 'threat' to one's group identity, we predicted more utilitarian (i.e., exclusion-oriented) and less restorative (i.e., inclusion-oriented) motives for punishing an ingroup compared to an outgroup offender (Brambilla *et al.*, 2013; Pagliaro *et al.*, 2013). In other words, we expected participants to try and restore their group image by assigning a punishment that involves a symbolic exclusion of the deviant member from their group. According to this reasoning, utilitarian punishments, which promote exclusion-oriented strategies such as incarceration, deportation, or even sanctioning the offender in public (see Carlsmith & Darley, 2008), may act as a tool that buffers threat effects of immoral doing. Restorative motives, in contrast, provide the opportunity to the transgressors to make up for their immorality and be empathized and forgiven instead of being treated with vengeance and affective blame (Lacey & Pickard, 2015; Zehr, 1997). Restoration is therefore more likely to be seen as a 'gentle' rather than firm response by third parties – and is unlikely to be as effective at helping group members manage the discomfort caused by the harm to the group's moral image (*Hypothesis 2*).

Further, we also aim to explore the moderating role of harm-doer's ingroup typicality in the relationship between group membership of the offender and motives for punishment. Considering Pinto *et al.* (2010) findings, according to which deviant typical ingroup members are derogated more and punished more harshly compared to deviant marginal (i.e., atypical) ingroup members or outgroup members, we propose that observers would display more exclusion-oriented punishments against ingroup offenders who are typical members of the ingroup. Typical ingroup offenders might be viewed as a higher threat to the group identity as compared to outgroup and atypical ingroup members because their immorality is generalized to the group and is already mirroring the group's moral values and norms. In short, the impact of typical ingroup members' behaviour is much stronger and might require more exclusionary punishments in order for the group to restore its image.

Hence, we hypothesize that typical ingroup as compared to atypical or outgroup offenders should be assigned more exclusion-oriented (i.e., utilitarian) and less proportionate (i.e., retributive) or inclusion-oriented (i.e., restorative) punishments. However, motives for punishment should follow the default pattern – retribution being more common than utilitarianism or restoration – when transgressors are atypical or outgroup members, as worry about image threat is less salient in these conditions (*Hypothesis 3*).

Further, we hypothesize that perceived ingroup identity threat should mediate the effect of offender's group membership and ingroup typicality on justice decisions. People experience increased identity threat when deviant members come from the ingroup rather than the outgroup (Okimoto & Wenzel, 2010; Rullo *et al.*, 2015; Sankaran *et al.*,

2017; Van der Toorn *et al.*, 2015) and when they are typical rather than atypical ingroup members (Levine & Moreland, 2002; Meeussen *et al.*, 2012). This is because group values and reputation are highly jeopardized in such cases. We claim that observers will assign stronger exclusion and less inclusion-oriented punishments to ingroup and especially to typical ingroup offenders through increased identity threat. Furthermore, we expect that ingroup threat will mediate the interaction effect between offender's group membership and ingroup typicality on justice reactions. Specifically, we hypothesize that when the offender is a typical ingroup member, observers will assign more exclusion-oriented than inclusion-oriented punishments through the experience of ingroup identity threat (*Hypothesis 4*).

Finally, we aim to investigate the moderating role of observers' ingroup identification on the effect of offenders' group membership and ingroup typicality on ingroup threat (mediator) and punishment motive. To the extent that group membership and group typicality are likely to be more self-conceptually important for high identifiers (Abrams *et al.*, 2000; Branscombe *et al.*, 1993; Castano *et al.*, 2002; Hutchison & Abrams, 2003), we would expect that offender's group membership and ingroup typicality to influence ingroup identity threat and in turn punishment motives particularly for high identifiers. We hypothesize that when the offender comes from the ingroup (as opposed to an outgroup) and is a typical (as opposed to atypical) member of the ingroup, high identifiers will experience stronger identity threat, which will in turn make them assign stronger exclusionary (i.e., utilitarian) punishments (*Hypothesis 5*). By contrast, we expect that ingroup threat and punishment motives will be less dependent on the offenders' group membership or ingroup typicality for low group identifiers.

In Study 1, we manipulated the group membership of the offender (ingroup vs. outgroup) and tested its effect on observers' motives for punishment. In Study 2, we additionally manipulated the offender's ingroup typicality (typical vs. atypical) and tested its role in the aforementioned relationship. In Study 3, we aimed to replicate Study 2 through a similar experimental design and we further assessed observers' ingroup identification (moderator) and ingroup identity threat (mediator). The offender's immoral act in all three studies was stealing money from a colleague.

STUDY 1

Methods

Participants

A total of 189 Dutch participants (93 females and 96 males; M_{age} = 37.6, SD = 13.25) living in the Netherlands took part in this study. We excluded two participants who failed the manipulation checks from further statistical analyses. The present sample gives 95% power to detect a medium effect size (f = .26).

Experimental design and procedure

We manipulated offender's group membership in vignettes (see Appendix S1). We recruited participants from crowded locations (i.e., metro and train stations) from several cities in the Netherlands in a paper-and-pencil research. Respondents were asked to read a scenario (in Dutch) which presented a case where a Dutch (ingroup) versus a Moroccan (outgroup) employee working in an international company stole a wallet that was left behind on a table after a meeting (see Appendix S1 for the vignette). Participants were

randomly assigned to one of the two experimental conditions. As a manipulation check, we asked participants to indicate the nationality of the harm-doer as either Dutch, Moroccan or Spanish. The study was anonymous and participation was voluntary. Participants were thanked and debriefed after filling in the questionnaire.

Measures

We used the 16-item motives for punishment scale recently used in (Fousiani & Demoulin, in press). The scale assessed the various motives for punishment, including (1) utilitarian motives and its sub-dimensions (private deterrence, public deterrence, and incapacitation[1]); (2) retributive motives; and (3) restorative/rehabilitative motives for punishment (1 = *absolutely disagree*, 7 = *absolutely agree*). Carlsmith (2008) and Kugler *et al.* (2013) relied on a similar scale. Cronbach's alpha was .76 for utilitarian motives, .74 for retributive motives, and .70 for restorative motives for punishment (see Appendix S1 for full scale).

Results

Table 1 displays the correlations between the variables. Participants' scores were submitted to a 2 (group membership of the offender: ingroup/outgroup) × 3 (motives for punishment: utilitarian, retributive, restorative) mixed analysis of variance (ANOVA) with group membership varying between participants and motives for punishment within them (see Table 2). The main effect of group membership was not significant, $F < 1$, *ns*. As expected, the motive main effect proved significant, $F_{2,187} = 62.31$, $p < .001$, $\eta^2 = .40$. In line with Hypothesis 1, a first contrast (C1) comparing retributive motives (+2) to utilitarian (−1) and restorative (−1) motives confirmed that participants reported higher retributive than utilitarian or restorative motives for punishing the offender, $t_{(188)} = 11.05$, $p < .001$, $\eta^2 = .34$. The second contrast (C2) comparing utilitarian (+1) and restorative motives (−1) was not significant, $t_{(188)} < 1$, *ns.*

More importantly, and as expected, the motive by group membership interaction was also significant, $F_{2,187} = 4.56$, $p < .05$, $\eta^2 = .05$. To probe this interaction, we first examined each of our two contrasts as a function of group membership. Whereas the C1 contrast by group membership was not significant, $t_{188} < 1$, *ns*, the C2 contrast interacted significantly with group membership, $t_{188} = -3.01$, $p = .003$, $\eta^2 = .04$. This interaction

Table 1. Pearson correlation coefficients between variables (Study 1)

	1	2	3
1. Utilitarian motives	1	.38***	.22**
2. Retributive motives		1	.60***
3. Restorative motives			1

**$p < .01$; ***$p < .001$.

[1] *The literature distinguishes between deterrent private, deterrent public, and incapacitative motives for punishment. All these motives aim at controlling harm-doer's future behavior and are therefore included under the umbrella of utilitarian motives for punishment (see Carlsmith & Darley, 2008). We did not refer to each of those dimensions separately as we did not expect any differences between them. Instead, we calculated a general mean, indicating utilitarian motives for punishment.*

Table 2. Means and standard deviations for motives for punishment (Study 1)

	Ingroup		Outgroup		Mean	
Motives for punishment	*M*	*SD*	*M*	*SD*	*M*	*SD*
Utilitarian	4.12	1.06	3.71	1.04	3.91	1.06
Retributive	5.08	1.24	5.11	1.26	5.10	1.24
Restorative	3.77	1.43	4.25	1.51	4.02	1.48

Note. All ratings were on 7-point scales ranging from 1 = *absolutely disagree* to 7 = *absolutely agree.*

revealed that observers assigned more utilitarian and less restorative punishments to the ingroup than the outgroup offender. These results confirmed Hypothesis 2 (for the means and standard deviations, see Table 2). We also examined the impact of group membership for each motive. As predicted, participants reported stronger utilitarian motives for punishing an ingroup than an outgroup offender, $t_{188} = 2.70$, $p = .008$. Group membership significantly affected restorative motives in the opposite direction, with participants reporting a stronger desire to restoratively punish the outgroup versus ingroup member $t_{188} = -2.26$, $p = .025$. Finally, there was no effect of group membership on retributive motives, $t_{188} < 1$, *ns*.[2]

Discussion

This study aimed to examine the motives of people when they assign punishments to ingroup and outgroup immoral doers. Because people display an intuitive preference for proportionate rather than extreme sanctions (Wenzel & Thielmann, 2006), observers were expected to display a generalized tendency to assign more retributive as compared to utilitarian or restorative motives for punishing a harm-doer. Our findings fully confirmed our first hypothesis replicating prior research.

Second, based on the BSE theory (Marques & Yzerbyt, 1988; Marques *et al.*, 1988), we further hypothesized that individuals would display different motives for punishment depending on the group membership of the transgressor. Specifically, we reasoned that they would prefer a utilitarian type of punishment for ingroup (*vs.* outgroup) members, as utilitarian punishment involves a symbolic exclusion that acts as a tool to buffer the threat to the ingroup moral image. In contrast, we reasoned they should shy away from restorative punishment for ingroup (vs. outgroup members), as restorative punishment is inclusion-oriented and is therefore likely to be seen as less effective at restoring the harmed image of their group. In line with this second hypothesis, participants assigned stronger utilitarian and less restorative punishments to the ingroup compared to an outgroup transgressor. Unexpectedly, findings revealed that observers preferred restorative punishments for outgroup offenders rather than the default pattern of retribution. Of note, restorative punishments, similarly with utilitarian, also aim to control offenders' future behaviour (Van Prooijen, 2018), but in a more educational manner as compared to utilitarian punishments which are exclusionary and intolerant.

[2] *We also conducted an additional study (*N = *66 participants from Germany,* M_{age} = *31.8,* SD = *13.18). Results replicated Study 1 and further showed that observers perceive ingroup offenders as less typical of the ingroup (i.e., another means to deflect moral blame from the ingroup). However, given the small size of the sample (and the associated low statistical power), we did not include this study in the main text. We report this study in the Appendix S1.*

STUDY 2

Methods

Participants

Our initial sample consisted of 213 participants. After excluding the non-German participants and participants with multiple identities (n = 10) as well as the participants who failed the manipulation checks (n = 81), the final sample consisted of 122 German participants, employees from different companies in Germany (M_{age} = 38.28, SD = 49.61). The present sample provided 95% power to detect a medium effect size (f = .37).

Experimental design and procedure

The study employed a 2 (group membership of the offender: ingroup vs. outgroup) × 2 (ingroup typicality: offender typical vs. atypical of the ingroup) between-participants design. Participants were randomly assigned to one of the conditions. We manipulated group membership of the offender by means of vignettes. Depending on condition, participants saw a picture story of either an ingroup (German) or an outgroup (Turk) employee who stole a wallet from a colleague's bag at work. Since participants were Germans, the ingroup employee was presented as German, while the outgroup employee was presented as Turk (see Appendix S1 for the vignettes and the pictures used in this study). Under each picture was a short text that provided general information about the company, the harm-doer, and the immoral act/stealing.

To manipulate ingroup typicality, we presented additional information about the offender at the very end of the picture story. Specifically, in the typical condition participants read: *'Many aspects of Murat Yildiz's/Dirk Müller's attitude and behavior, such as his ideas, humor, and everyday interaction with others are very similar to the general attitude and behavior of most Germans. Most people see him as sharing the average German's worldview and value system and therefore Murat Yildiz/Dirk Müller is definitely perceived by his friends and colleagues as a typically German person'. In the atypical condition,* participants read: *'Many aspects of Murat Yildiz's/ Dirk Müller's attitude and behavior, such as his ideas, humor, and everyday interactions with others deviate a lot from the general attitude and behavior of most Germans. Most people see him as having a totally different worldview and value system compared to Germans and therefore Murat Yildiz/Dirk Müller is not perceived by his friends and colleagues as a typically German person'.*

Manipulation checks for group membership of the offender and typicality to the ingroup followed the picture story. Regarding group membership, because the story involved a German or a Turkish transgressor, we asked participants to indicate the nationality of the offender as German, Turkish, or Chinese. Regarding typicality to the ingroup, we provided participants with a two-choice question asking to indicate whether the offender of the vignette was or was not presented as a typical German person.

Respondents participated voluntarily and anonymously in an online survey via Qualtrics. The survey was in participants' native language (German).

Measures

We used the same 16-item scale as in Study 1 for the assessment of the various motives for punishment of the harm-doer (1 = *absolutely disagree*, 7 = *absolutely agree*). We

translated the scale into German, adjusting it to the needs of this study and the current vignettes. Cronbach's α's were .74, .86, and .80 for retributive, utilitarian, and restorative motives, respectively.

Results

Table 3 shows correlations between the variables. These correlations yielded a similar pattern as in Study 1, suggesting that our key constructs (e.g., retributive, utilitarian, restorative motives) are related but distinct constructs. We submitted participants' scores to a 2 (group membership of the offender: ingroup/outgroup) × 2 (typicality to the ingroup: typical/atypical) × 3 (motives for punishment: utilitarian, retributive, restorative) mixed ANOVA with group membership and typicality to the ingroup varying between participants and motives for punishment within them (see Table 4).

The main effects of group membership and typicality to the ingroup were not significant *F*s <1, *ns*. As expected, the punishment motive main effect was significant, $F_{2,111} = 31.80$, $p < .001$, $\eta^2 = .36$. Consistent with Hypothesis 1, a first contrast (C1) comparing retributive motives (+2) on the one hand and utilitarian and restorative motives on the other (−1) confirmed that participants reported higher retributive ($M = 4.95$, $SD = 1.48$) than utilitarian ($M = 3.57$, $SD = 1.38$) or restorative ($M = 4.24$, $SD = 1.68$) motives for punishing the offender, $t_{112} = 7.29$, $p < .001$, $\eta^2 = .32$. The second contrast (C2) comparing utilitarian (+1) and restorative motives (−1) was also significant, $t_{112} = 2.73$, $p < 01$, $\eta^2 = .06$, showing that participants reported higher restorative than utilitarian motives for punishment.

The punishment motive by group membership interaction was significant, $F_{2,111} = 4.22$, $p = .02$, $\eta^2 = .07$. To investigate this interaction, we first examined each of our two contrasts as a function of group membership. The C1 contrast by group membership was significant, $t_{112} = 2.84$, $p = .005$, $\eta^2 = .07$. Participants reported stronger retributive ($M = 5.16$, $SD = 1.49$) than utilitarian ($M = 3.31$, $SD = 1.47$) or restorative ($M = 4.11$, $SD = 1.71$) motives for punishing the outgroup, $t_{50} = 6.63$, $p < .001$, $\eta^2 = .47$. Albeit less pronounced, a similar pattern of results emerged for ingroup offender ($M_{retr} = 4.78$, $SD = 1.45$, $M_{util} = 3.77$, $SD = 1.28$, M_{restor} =c4.34, $SD = 1.66$), $t_{62} = 3.42$, $p = .001$, $\eta^2 = .16$. The C2 contrast did not interact significantly with group membership, $t_{112} < 1$, *ns*. We also looked at the impact of group membership of the offender separately for each motive. As expected, participants reported marginally significantly stronger utilitarian motives for punishing an ingroup ($M = 3.78$, $SD = 1.28$) rather than an outgroup ($M = 3.31$, $SD = 1.47$) offender, $t_{114} = 1.80$, $p = .07$, replicating the findings of Study 1. Group membership of the offender did not affect significantly either retributive or restorative motives, $t_{114} < 1$, *ns*.

Table 3. Pearson correlations coefficients between variables (Study 2)

	1	2	3
1. Utilitarian motives	1	.05	−.24**
2. Retributive motives		1	.14
3. Restorative motives			1

Note. **$p < .01$.

Table 4. Means and standard deviations for motives for punishment (Study 2)

	Typical				Atypical			
	Ingroup		Outgroup		Ingroup		Outgroup	
Motives for punishment	M	*SD*	M	*SD*	M	*SD*	M	*SD*
1. Utilitarian	4.27	1.42	3.23	1.63	3.40	1.04	3.36	1.38
2. Retributive	4.55	1.34	5.21	1.52	4.96	1.53	5.13	1.49
3. Restorative	4.61	1.58	3.77	1.67	4.13	1.72	4.31	1.73

Note. All ratings were on 7-point scales ranging from 1 = *absolutely disagree* to 7 = *absolutely agree*.

The punishment motive by typicality to the ingroup interaction was not significant, $F < 1$, *ns.*

Finally, and importantly, the three-way interaction between punishment motives, group membership of the offender, and typicality to the ingroup was significant $F_{2,111} = 3.15$, $p = .047$, $\eta^2 = .054$ (see Table 4). To unpack this interaction, we examined the group membership by punishment motives interaction at each level of typicality. The punishment motive by group membership interaction was not significant when the offender was atypical of the ingroup, $F < 1$, *ns.* In sharp contrast, and as expected, the interaction was significant when the offender was typical of the ingroup, $F_{2,44} = 5.85, p = .006, \eta^2 = .21$. Probing this interaction revealed that the C1 contrast by group membership was significant, $t_{45} = 3.46$, $p = .001$, $\eta^2 = .21$, in that participants reported stronger retributive motives compared to the other two motives when the offender was in fact an outgroup member but nevertheless typical of the ingroup. Unexpectedly, the second contrast (C2) opposing utilitarian (+1) and restorative motives (−1) did not interact significantly with group membership, $t_{45} < ns$ (means and standard deviations are presented in Table 4).

We also examined the three-way interaction by looking at the group membership of the offender by typicality interaction separately for each punishment motive. The interaction was marginally significant for utilitarian motives, $F_{1,112} = 3.74$, $p = .056$, $\eta^2 = .03$, showing that participants opted for more utilitarian punishment towards a typical versus atypical ingroup offender. The mean difference was not significant for outgroup offenders. Finally, the interaction was not significant for either retributive or restorative punishment motives, Fs <1, *ns* (means and standard deviations are presented in Table 4).

Discussion

In this study, we aimed to replicate Study 1 and experimentally investigate the moderating role of offender's typicality to the ingroup in the relationship between group membership and punishment motives. As expected, findings showed that people in general display a stronger preference for retributive punishments, confirming Hypothesis 1. Hypothesis 2 was partly confirmed: In line with our predictions, people preferred more exclusionary (i.e., utilitarian) punishment for an ingroup as compared to an outgroup offender, in accordance with the BSE theory (Marques & Yzerbyt, 1988; Marques *et al.*, 1988). However, group membership had no effect on either retributive or inclusionary (i.e., restorative) punishment motive. Similarly, Hypothesis 3 was partly confirmed. Although

people engaged in more utilitarian punishment towards an ingroup member who was typical as opposed to atypical of the ingroup, ingroup typicality had no effect on either retributive or restorative punishment against ingroup offenders.

Finally, an interesting pattern of results emerged with regard to outgroup offenders who displayed traits typical of the ingroup. These offenders were assigned stronger retributive than restorative or utilitarian punishments. This suggests that when offenders come from an outgroup, they are not perceived as a threat even if they share characteristics that are typical of the ingroup. Presumably, this is because when an individual does not belong to the ingroup (even if they have some traits typical of it) observers no longer find themselves compelled to protect their group reputation. Therefore, they can default to the more common intuitive (i.e., retributive) motives for punishment.

STUDY 3

In Study 3, we aimed to achieve a better understanding of the psychological mechanisms that motivate individuals to treat ingroup offenders as black sheep and thus assign to them exclusionary (i.e., utilitarian) punishments. To this end, we further investigated the role of identity threat as mediating the relationship between group membership and ingroup typicality of the offender and observers' justice reactions. People, when confronted with immoral acts enacted by ingroup perpetrators or perpetrators who are typical of the ingroup, experience increased threat to their group image (Meeussen *et al.*, 2012; Piff *et al.*, 2012; Rullo *et al.*, 2015; Van der Toorn *et al.*, 2015). In line with this literature, we hypothesized that perceived ingroup identity threat should mediate the effect of offender's group membership and ingroup typicality on justice decisions. Additionally, we conjectured that this effect might be stronger for high identifiers (as opposed to low identifiers), because group membership and group typicality would be more important for high than for low identifiers (Abrams *et al.*, 2000; Branscombe *et al.*, 1993; Castano *et al.*, 2002; Hutchison & Abrams, 2003).

Methods

Participants

A total of 307 British participants initially took part in the study. After excluding those who failed the manipulation checks, the final sample consisted of 292 (180 females and 112 males; M_{age} = 37.13, SD = 10.97) participants living in the United Kingdom. An *a priori* power analysis revealed that, using our design, 270 participants were required in order to achieve 80% power to detect a medium effect size (f = .25). Participants were recruited via Prolific academic and were paid £ 0.80 (€.90) for their participation. The study was programmed in Qualtrics such that each IP address could participate only once.

Experimental design and procedure

In order to separate the assessment of ingroup identification from the experimental manipulations, respondents learned they would take part to two separate studies. First, we asked participants to fill in an ingroup identification scale. The experimental design and the assessment of dependent variables followed. The study relied on a 2 (group membership of the offender: ingroup vs. outgroup) × 2 (ingroup typicality: offender typical vs. atypical of the ingroup) between-participants design. Participants were

randomly assigned to one of the conditions. As in Study 2, we manipulated group membership of the offender by means of vignettes. Depending on condition, participants saw a picture story of either an ingroup (British) or an outgroup (Indian) employee who stole a wallet from a colleague's bag at work. As with Study 2, we manipulated ingroup typicality by presenting additional information about the offender at the very end of the picture story (see Appendix S1 for the vignettes and the pictures used in this study).

Manipulation checks for group membership and typicality followed the picture story. Manipulation checks were similar to the ones use in Study 2, albeit adjusted to the content of the vignettes of the current study. The survey was conducted in participants' native language (English).

Measures

Ingroup identification

We assessed participants' identification with the ingroup with 10 items (1 = *absolutely disagree*, 7 = *absolutely agree*) taken from the 'identity' factor of the collective self-esteem scale (Luhtanen & Crocker, 1992) and the identification scales used in Verkuyten (2005) and Stephan *et al.* (2002). Some example items are 'Overall, being British has very little to do with how I feel about myself' 'My British identity is an important reflection of who I am'. Cronbach's alpha for this scale was .94.

Ingroup identity threat

We assessed participants' ingroup identity threat with Duckitt's (2006) scale. The scale consisted of eight items using a 7-point scale (1 = *absolutely disagree*, 7 = *absolutely agree*). High threat items described employees like the one presented in the vignette as undermining important social values, norms, and traditions as well as threatening security and stability in society (e.g., 'Employees like the one presented in the article. . . seem to want to destroy or harm what is good in our society'). Low threat items described employee of the vignette as making British society safer, stronger, and more united (e.g., '. . .help to make British society stronger and more united'). Cronbach's alpha for this scale was .86.

Motives for punishment

A revised 16-item scale (Fousiani & Demoulin, in press) similar to the one used in Studies 1 and 2 allowed the assessment of the various motives for punishment of the harm-doer (1 = *absolutely disagree*, 7 = *absolutely agree*). We translated the scale into English and adjusted it to the needs of this study and the current vignettes. Cronbach's α's were .85, .89, and .71 for retributive, utilitarian, and restorative motives, respectively.

We report all scales in the Appendix S1.

Results

Table 5 shows the correlations between the variables. Again, these correlations yielded a pattern similar to the one found in Studies 1 and 2, suggesting that our key constructs (e.g., retributive, utilitarian, and restorative motives) are related but distinct constructs. Participants' scores were submitted to a 2 (group membership of the offender: ingroup/outgroup) $\times$ 2 (ingroup typicality: typical/atypical) $\times$ 3 (motives for punishment:

Table 5. Pearson correlation coefficients between variables (Study 3)

	1	2	3	4	5
1. Utilitarian motives	1	.37***	−.24**	.17**	.38***
2. Retributive motives		1	.01	−.07	.11
3. Restorative motives			1	−.01	−.08
4. Ingroup identification				1	.37***
5. Ingroup identity threat					1

**$p < .01$; ***$p < .001$.

utilitarian, retributive, restorative) mixed ANOVA with group membership of the offender and ingroup typicality varying between participants and motives for punishment within them (see Table 6).

The main effects of group membership and typicality on motives were not significant *F*s < 1, *ns.* As expected, the punishment motive main effect was significant, $F_{2,287} = 206.27$, $p < .001$, $\eta^2 = .59$. Consistent with Hypothesis 1, a first contrast (C1) comparing retributive motives (+2) on the one hand and utilitarian and restorative motives on the other (−1) confirmed that participants reported higher retributive ($M = 5.88$, $SD = 1.06$) than utilitarian ($M = 4.91$, $SD = 1.26$) or restorative ($M = 4.17$, $SD = 1.38$) motives for punishing the offender, $t_{288} = 20.35$, $p < .001$. The second contrast (C2) comparing utilitarian (+1) and restorative motives (−1) was also significant, $t_{288} = 5.95$, $p < .001$, showing that participants reported higher utilitarian than restorative motives for punishment.

The punishment motive by group membership interaction was significant, $F_{2,287} = 3.13$, $p = .04$, $\eta^2 = .02$. To probe this interaction, we first examined each of our two contrasts as a function of group membership. The C1 contrast by group membership was marginally significant, $t_{288} = 1.79$, $p = .07$, $\eta^2 = .01$, revealing a stronger preference for assigning a retributive rather than a utilitarian or restorative punishment to an ingroup offender ($M_{retr} = 5.87$, $SD = 1.02$, $M_{util} = 5.07$, $SD = 1.18$, $M_{restor} = 4.19$, $SD = 1.31$) $t_{152} = 15.49$, $p < .001$, $\eta^2 = .62$. Similarly, participants displayed stronger retributive as opposed to utilitarian or restorative motives for punishing an outgroup offender ($M_{retr} = 5.90$, $SD = 1.11$, $M_{util} = 4.74$, $SD = 1.33$, $M_{restor} = 4.14$, $SD = 1.46$) $t_{136} = 13.52$, $p < .001$, $\eta^2 = .57$. Contrary to our hypothesis (but consistent with Studies 1 and 2), the C2 contrast did not interact significantly with group membership, $t_{288} < 1$, *ns.* We also looked at the impact of group membership separately for each motive. As expected, participants reported significantly stronger utilitarian motives for punishing an ingroup ($M = 5.07$, $SD = 1.18$) rather than an outgroup ($M = 4.74$, $SD = 1.33$) offender, $t_{290} = 2.26$, $p = .02$, replicating Studies 1 and 2. Again, group membership did not affect significantly either retributive or restorative motives, $t_{290} < 1$, *ns*.

The punishment motive by ingroup typicality interaction was not significant, $F < 1$, *ns.* Unexpectedly, the three-way interaction effect between punishment motives, group membership of the offender, and ingroup typicality failed to reach significance, $F < 1$, *ns*.

Next, we conducted a mediation analysis with group membership of the offender as the independent variable (effect coded −1 = ingroup; 1 = outgroup), utilitarian motives for punishment as the dependent variable, and perceived ingroup threat as mediator. The total effect of group membership on utilitarian motives was negative and significant:

Table 6. Means and standard deviations for motives for punishment (Study 3)

	Ingroup		Outgroup		Mean	
Motives for punishment	*M*	*SD*	*M*	*SD*	*M*	*SD*
Utilitarian	5.07	1.18	4.74	1.33	4.91	1.26
Retributive	5.87	1.02	5.90	1.11	5.88	1.06
Restorative	4.19	1.32	4.14	1.46	4.17	1.38
Ingroup identity threat	5.05	1.08	4.18	1.28	4.64	1.25

Note. All ratings were on 7-point scales ranging from 1 = *absolutely disagree* to 7 = *absolutely agree*.

Observers assigned stronger utilitarian motives when the offender was an ingroup as opposed to outgroup member. When we included perceived threat towards the ingroup as a mediator in the model, both paths comprising the indirect effect proved significant and fully mediated the effect of condition (Yzerbyt, Muller, Batailler, & Judd, 2018, see Table 7 for the relevant statistics). These results support Hypothesis 4.

Finally, we tested whether the mediated effect of group membership on punishment motive through ingroup threat varies as a function of observers' ingroup identification. We ran a moderated mediation analysis relying on the sequential steps advocated by Muller, Judd, and Yzerbyt (2005; see also Hayes, 2013). Prior to these analyses, we mean-centred both ingroup identification (moderator) and ingroup threat (mediator). First, a model with group membership, ingroup identification, and their interaction as predictors and utilitarian motives as the criterion revealed a significant effect of group membership, $b = -0.16$, $t_{288} = 2.25$, $p < .03$, and of ingroup identification, $b = 0.16$, $t_{288} = 2.76$, $p < .001$, but no significant interaction, $b = 0.09$, $t_{288} = 1.15$, $p > .13$. Interestingly, the same model using threat as the criterion confirmed the impact of group membership, $b = -0.43$, $t_{288} = 6.74$, $p < .001$, and ingroup identification, $b = 0.35$, $t_{288} = 6.88$, $p < .001$. This time, and in line with our intuitions, the two-way interaction came close to significance, $b = 0.09$, $t_{288} = 1.74$, $p = .08$, showing that the impact of group membership on threat tended to be stronger for high identifiers than for low identifiers. As our final step, we examine a model that included the same three predictors along with threat and the interaction between threat and ingroup identification to predict utilitarian motives. Only the mediator, that is, threat, proved significant, $b = 0.35$, $t_{286} = 5.44$, $p < .001$. Contrary to Hypothesis 5, thus, the data fail to confirm the viability of a moderated mediation involving ingroup identification.

Discussion

In this study, we aimed to replicate Studies 1 and 2 and further investigate whether ingroup threat mediates the relationship between group membership of the offender and typicality of the offender to the ingroup, and punishment motive. Moreover, we aimed to test whether the above mediational effect varies as a function of the observers' ingroup identification.

First, the findings of this study replicated the findings of Studies 1 and 2 showing that when the offender comes from the ingroup (as opposed to an outgroup), perceivers assign more exclusion-oriented punishment (i.e., utilitarian). Further, we hypothesized that when the offender comes from the ingroup, perceivers would experience higher ingroup

Table 7. Mediation results with perceived threat as mediator (Study 3)

Effects of group membership of the offender on	Total effect	Direct effect (c′)	Unstandardized paths		Indirect effect			Ratio of indirect to total effect
			a	*b*	Estimate	95% CI lower	95% CI upper	
Utilitarian motives	−.17 (.07)*	−.0001 (.07)	−.43 (.07)***	.38 (.06)***	−.16 (.04)	−.26	−.10	.99

Notes. Standard errors in parentheses (bootstrap standard errors for the indirect effect estimate); CI: percentile bootstrap confidence interval; paths *a* and *b* correspond to the prediction coefficients of the independent variable to the mediator (path *a*) and of the mediator to the dependent variable (path *b*).
*$p < .05$; ***$p < .001$.

identity threat, which would in turn lead to a stronger preference for exclusionary punishments (i.e., utilitarian) against the offender. The present findings supported Hypothesis 4, revealing a mediating effect of ingroup threat on the relationship between offender's group membership and utilitarian punishment motive. Unexpectedly, and in contrast to Study 2, the effect of ingroup typicality on punishment motive was not significant. Finally, we hypothesized that the effect of group membership and typicality on punishment motive through the experience of ingroup threat would be stronger for high identifiers. Although there was a significant main effect of ingroup identification on both perceived threat and utilitarian punishment – with high identifiers generally experiencing an increased sense of threat but also displaying more utilitarian punishment motive – there was no interaction between ingroup identification and group membership in the prediction of utilitarian motives via threat. Of note, the interaction of group membership and ingroup identification on identity threat was marginally significant. As predicted, findings revealed that high (as opposed to low) identifiers experience stronger identity threat when the offender comes from the ingroup. Finally, ingroup typicality did not influence punishment motive. Hypothesis 5 was therefore not supported.

Overall, these findings are in line with our argument that people display a strong preference for utilitarian punishments (but not restorative or retributive) when sanctioning ingroup offenders. Utilitarian punishments can be viewed as exclusion-oriented and may thus signal a symbolic exclusion of ingroup deviants aiming at restoring the harmed group image. Additionally, our results suggest the perceived ingroup identity threat is a key psychological mechanism that helps to explain this effect.

GENERAL DISCUSSION

People not only have stereotypes about a wide variety of groups, but they also care about how others perceive their group (Yzerbyt & Demoulin, 2010). Immorality is considered a great threat to the image of one's group (Brambilla *et al.*, 2013; Pagliaro *et al.*, 2013). Individuals apply several strategies against immoral doers, including rejection of transgressors, denial of the severity of a transgression, or calling for transgressors' remorse in order to protect their group reputation and avoid contamination (Van Leeuwen *et al.*, 2010). Ample research has systematically shown people's tendency to derogate deviant ingroup members as a means to protect the group from the threat that they pose to their identity (Abrams, Marques, Randsley de Moura, Hutchison, & Bown, 2006; Abrams, Rutland, Ferrell, & Pelletier, 2008; Marques & Paez, 1994; Marques & Yzerbyt, 1988; Marques *et al.*, 1988). Limited research exists, however, on how information about a harm-doers' group membership affects justice-related decisions (Braun & Gollwitzer, 2012; Brown *et al.*, 2008; Gollwitzer & Keller, 2010; Gollwitzer & Van Prooijen, 2016; Van Prooijen & Lam, 2007), and no prior research exists on the motives underlying individuals' assignment of different types of punishment to ingroup versus outgroup transgressors. In Study 1, we tested the effect of a transgressor's group membership on observers' motives for punishment. In Study 2, we examined the moderating role that a transgressor's ingroup typicality might play in the relationship between the transgressor's group membership and observers' motives for punishment. Finally, in Study 3 we investigated the mediating effect of ingroup identity threat as well as the moderating effect of observers' ingroup identification in the above relationships.

First, given that people's intuitions of justice tend by default to be more retributive than utilitarian or restorative (Carlsmith, 2008; Carlsmith *et al.*, 2002; Darley & Pittman,

2003), we hypothesized that more retributive as compared to utilitarian or restorative motives for punishment would generally be displayed. Results fully corroborated this hypothesis. Every single study showed that retribution prevails over the other two types of punishment.

Second, we hypothesized that people would display more utilitarian and less restorative motives for punishing an immoral perpetrator who comes from the ingroup rather than from an outgroup. Immoral conduct committed by a fellow group member represents a threat to one's group identity and jeopardizes the ingroup's moral standing (Van der Toorn *et al.*, 2015). Assigning utilitarian punishments against a 'deviant' ingroup member might – by virtue of utilitarian punishments' relative harshness and exclusionary nature – indicate a symbolic distancing of this member from the group, thereby contributing to the restoration of the group's reputation. Study 1 provided full support to our hypothesis, showing that people administer more utilitarian than restorative motives for punishment to ingroup than outgroup harm-doers. Study 2 partly replicated these findings showing that people display stronger exclusionary punishments against ingroup as opposed to outgroup offenders (although here we observed no effect of group membership on either retributive or restorative punishment). Study 3 again confirmed the link between ingroup membership and greater utilitarian punishment motives. These findings are consistent with prior research on the particular importance individuals place on the morality of the ingroup. For instance, Van der Lee, Ellemers, Scheepers, and Rutjens (2017) found that immoral but not incompetent individuals were perceived as more different from the ingroup and were more likely to be rejected. In line with our own theorizing, these authors also found that threat to the ingroup mediated the rejection of immoral members.

In Study 2, we further explored the role of ingroup typicality and its moderating effects on the relationship between offender's group membership and motives for punishment. We hypothesized that when the ingroup offender shares similar values with the rest fellow members and therefore comes across as a representative/typical member of his group, perceived threat to one's group identity would be higher. We reasoned that observers would thus assign more exclusion-oriented (i.e., utilitarian) than retributive or inclusionary (i.e., restorative) punishments against typical ingroup offenders, in order to demonstrate to others that such behaviours are not tolerated by the ingroup. Results partly confirmed our hypothesis showing that observers preferred to administer utilitarian punishments to typical as compared to atypical ingroup offenders. Our findings are in line with prior research showing that deviant ingroup typical members are derogated more as opposed to deviant atypical members (Pinto *et al.*, 2010). Typical ingroup members' immorality reflects heavily on the values and standing of the entire group (Castano *et al.*, 2002; Meeussen *et al.*, 2012). Given this threat to the group's identity, observers treat them as black sheep via exclusion-oriented (i.e., utilitarian) punishments that distance the ingroup from the immoral behaviour. Unexpectedly, ingroup typicality had no effect on retributive or restorative punishments against ingroup offenders.

Finally, Study 3 replicated Studies 1 and 2 and further investigated the mediating effect of ingroup threat in the relationship between group membership and ingroup typicality of the offender and punishment motive. In line with our predictions, we found that an observer experiences greater threat when faced with an ingroup rather than an outgroup offender, and in turn assigns more utilitarian punishment to this offender. This finding is in line with prior research which suggests that people experience increased identity threat when deviant members come from the ingroup rather than the outgroup (Okimoto & Wenzel, 2010; Rullo *et al.*, 2015; Sankaran *et al.*, 2017; Van der Toorn *et al.*, 2015). To the

best of our knowledge, this research is the first to show that people display stronger preference for (specifically) exclusionary punishments against ingroup offenders through the experience of increased ingroup threat.

Furthermore, although high ingroup identifiers indicated having experienced higher identity threat and displayed stronger utilitarian punishments against an offender, we observed no support for our prediction that identification with the ingroup would moderate the effect of group membership on perceived threat and, in turn, on punishment motive. Moreover, ingroup typicality did not influence punishment motive in Study 3 (as opposed to Study 2). These results are inconsistent with prior findings which point out the particular importance that dimensions such as group membership and group typicality have for high (as opposed to low) identifiers (Abrams *et al.*, 2000; Branscombe *et al.*, 1993; Castano *et al.*, 2002; Hutchison & Abrams, 2003). Future research might replicate these findings using alternative measures for the assessment of identification with the ingroup.

Indeed, ingroup identification is typically seen as a multi-dimensional construct (Leach *et al.*, 2008) and it is likely that specific dimensions not examined here might be particularly important for justice-related reactions when ingroup (vs. outgroup) members are involved. For instance, 'ingroup homogeneity', a main factor of ingroup identification (Leach *et al.*, 2008), might be especially relevant to black sheep reactions against deviant ingroup members. Prior research, for example, has indicated that perceived entitativity (i.e., the degree to which a group appears to be a unified social entity) which is tightly related to the concept of ingroup homogeneity (Hamilton, Sherman, & Rodgers, 2004) may help explain ingroup derogation (Lewis & Sherman, 2010; Yzerbyt, Castano, Leyens, & Paladino, 2000).

Future work might also consider reasons for the inconsistencies in the interaction between group membership and ingroup typicality on punishment motive. Whereas the effect of the offender's group membership on punishment motive was robust across all three studies, the interaction effect between group membership and typicality was significant in Study 2 but not in Study 3 (ingroup typicality was not manipulated in Study 1). We also observed that the overall preference of observers for each punishment motive varied somewhat across the three studies. What might account for these differences? One explanation could be that the proportion of ingroup and outgroup members that live in the three different countries in which the studies were conducted varies. For example, although in all three countries a large number of the total population is comprised of outgroup members, the United Kingdom (Study 3) is more homogenous (87.1% of the total population are White British; https://www.indexmundi.com/united_kingdom/) as compared to the Netherlands (Study 1; 77.4% are Dutch; https://www.indexmundi.com/netherlands/) and Germany (Study 2; 79% are German; https://www.indexmundi.com/germany/). One might speculate that differences in the proportion of minority versus majority members may influence the level of prejudice towards minority members, the level of identification with the ingroup, or even the level of perceived threat by the immorality (see Barlow, Hornsey, Thai, Sengupta, & Sibley, 2013; Pettigrew, Wagner, & Christ, 2010). In our study in particular, British participants displayed stronger utilitarian punishment motive towards ingroup deviants as compared to Dutch or German participants, a finding that can be partly explained by the higher value that more homogenous groups may place on the protection of their ingroup reputation and image. Put differently, more homogenous countries may identify with their ingroup more strongly and thus experience increased threat when immoralities come from the ingroup (see also Lee & Ottati, 1995). This might lead them in turn to treat ingroup deviants as black sheep.

Limitations and future directions

Despite its contributions, our work has some theoretical and methodological limitations worth acknowledging. First, in Study 2, (but not in Study 3) a large number of participants failed the manipulation checks and were thus excluded from further analyses. This might be because this was an online study, in contrast to Study 1's paper-and-pencil design. A second limitation is our use of self-report instruments. It remains to be seen whether the nature of actual punishment behaviour would correspond to what participants report. In addition, although the development of the motives for punishment scale was inspired by prior well-established scales (Carlsmith, 2008; Kugler *et al.*, 2013), this measure was recently developed for the aims of the present research. Future research should focus on the development of validated scales for the assessment of motives for punishment. It should also be noted that this study focused on observers' punitive intentions rather than punitive behaviour against the transgressors. Future research using behavioural measures in order to replicate these effects would be valuable.

Moreover, future research should explore the differential role of sub-dimensions of utilitarian motives (see Carlsmith & Darley, 2008; Nagin, 1998; Van Prooijen, 2018) in black sheep situations. Deterrence, on the one hand, assumes that the offender is a rational person who has designed and deliberately committed an offence, and calculated the expected benefits in relation to the expected costs (e.g., likelihood and severity of punishment). Incapacitation, on the other hand, assumes that the offender is unable to think or act rationally and needs to be restrained into a legal quarantine so that the prevalence of crime is reduced. It would be worthwhile investigating which of these types of punishment is preferred against offenders who are seen as a physical threat who jeopardizes the reputation of the group. This way we could make conclusions about whether 'threatening offenders' are perceived as rational persons and how this may affect the way they are punished. Finally, recent literature includes restorative punishments under the umbrella of utilitarian punitive motives together with deterrence and incapacitation. In fact, all three punishments (deterrent, incapacitative, and restorative) aim to control offenders' future behaviour and thus share some sort of 'utility' (Van Prooijen, 2018). Future studies should take this into account for a more nuanced and accurate picture of the utilitarian approach to the protection of group image.

Along the same lines, it should be noted that the concept of utilitarianism in moral decisions itself has multiple facets. As outlined in the two-dimensional model of utilitarianism (Kahane *et al.*, 2018), utilitarian decision-making involves at least two psychological dissociable and independently important aspects, namely 'instrumental harm' and 'impartial beneficence' (see also Everett *et al.*, 2018). The symbolic exclusion of ingroup deviants might be explained through either of these two dimensions of utilitarianism (or both). On the one hand, members might instrumentally harm a 'black sheep' in order to achieve the broader of restoring the ingroup's reputation and image. On the other hand, they show a lack of partisan partiality by penalizing ingroup members just as harshly as (or more harshly) than they do outgroups.

Finally, in the present study we manipulated transgression via theft. Theft is a universally unacceptable behaviour that most cultures reject and punish. Still, it is only one type of moral transgression, and generalizing to other types would be valuable. Moreover, future research can move beyond considering moral acts that are universally considered violations across groups and cultures (such as fairness, or harm/care; Haidt, 2001) by considering how the violation of *group-specific moral norms* (see Ellemers, Pagliaro, & Barreto, 2014) might affect reputation-protective group reactions.

Conclusion

Taken together, these findings shed light on the role of intergroup phenomena on justice reactions. Justice decisions are not simply neutral judgements, but rather appear to serve functional goals – serving to protect against symbolic and actual threats to one's social identity. By punishing ingroup versus outgroup transgressions in more utilitarian (vs. retributive or restorative) ways, individuals reaffirm the values underpinning their social identities.

Compliance with ethical standards

This research involves human participants. All procedures performed in this study were in accordance with the ethical standards of the institutional research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Data availability statement

Data (dataset including scales used in this study) are available from the Open Science Framework at https://osf.io/scuzm/?view_only=1795bcc99c264c398cae36040c475477.

References

Abrams, D., Marques, J. M., Bown, N. J., & Henson, M. (2000). Pro-norm and anti-norm deviance. *Journal of Personality and Social Psychology*, *78*, 906–912.

Abrams, D., Marques, J. M., Randsley de Moura, G., Hutchison, P., & Bown, N. J. (2006). The maintenance of entitativity: A subjective group dynamics approach. In V. Y. Yzerbyt, C. M. Judd & O. Corneille (Eds.), *The psychology of group perception: Contributions to the study of homogeneity, entitatitivy, and essentialism*. Philadelphia, PA: Psychology Press.

Abrams, D., Palmer, S. B., Rutland, A., Cameron, L., & Van de Vyver, J. (2014). Evaluations of and reasoning about normative and deviant ingroup and outgroup members: Development of the black sheep effect. *Developmental Psychology*, *50*, 258–270. https://doi.org/10.1037/a0032461.

Abrams, D., Rutland, A., Ferrell, J. M., & Pelletier, J. (2008). Children's judgments of disloyal and immoral peer behavior: Subjective group dynamics in minimal intergroup contexts. *Child Development*, *79*, 444–461. https://doi.org/10.1111/j.1467-8624.2007.01135.x

Barlow, F. K., Hornsey, M. J., Thai, M., Sengupta, N. K., & Sibley, C. G. (2013). The wallpaper effect: The contact hypothesis fails for minority group members who live in areas with a high proportion of majority group members. *PLoS ONE*, *8*(12), e82228. https://doi.org/10.1371/journal.pone.0082228

Bentham, J. (1789). *An introduction to the principles of morals and legislation*. Oxford: Clarendon Press, 1907. https://doi.org/10.1093/actrade/9780198205166.book.1

Brambilla, M., Sacchi, S., Pagliaro, S., & Ellemers, N. (2013). Morality and intergroup relations: Threats to safety and group image predict the desire to interact with outgroup and ingroup members. *Journal of Experimental Social Psychology*, *49*, 811–821. https://doi.org/10.1016/j.jesp.2013.04.005

Branscombe, N. R., Doosje, B., & McGarty, C. (2002). Antecedents and consequences of group-based guilt. In D. M. Mackie & E. R. Smith (Eds.), *From prejudice to intergroup emotions: Differentiated reactions to social groups* (pp. 49–66). Philadelphia, PA: Psychology Press.

Branscombe, N. R., Ellemers, N., Spears, R., & Doosje, B. (1999). The context and content of social identity threat. In N. Ellemers, R. Spears & B. Doosje (Eds.), *Social identity: Context, commitment, content* (pp. 35–58). Oxford, UK: Blackwell.

Branscombe, N., Wann, D. L., Noel, J. G., & Coleman, J. (1993). In-group or out-group extremity: Importance of the threatened social identity. *Personality and Social Psychology Bulletin*, *17*, 381–388. https://doi.org/10.1177/0146167293194003

Braun, J., & Gollwitzer, M. (2012). Leniency for out-group offenders. *European Journal of Social Psychology*, *42*, 883–892. https://doi.org/10.1002/ejsp.1908

Brown, R., González, R., Zagefka, H., Manzi, J., & Čehajić, S. (2008). Nuestra Culpa: Collective guilt as a predictor for reparation for historical wrong-doing. *Journal of Personality and Social Psychology*, *94*, 75–90. https://doi.org/10.1037/0022-3514.94.1.75

Carlsmith, K. M. (2008). On justifying punishment: The discrepancy between words and actions. *Social Justice Research*, *21*, 119–137. https://doi.org/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. https://doi.org/10.1037/0022-3514.83.2.284

Castano, E., Paladino, M.-P., Coull, A., & Yzerbyt, V. (2002). Protecting the ingroup stereotype: Ingroup identification and the management of deviant ingroup members. *British Journal of Social Psychology*, *41*, 365–385. https://doi.org/10.1348/014466602760344269

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. https://doi.org/10.1023/A:100555220

Darley, J. M., & Pittman, T. S. (2003). The psychology of compensatory and retributive justice. *Personality and Social Psychology Review*, *7*, 324–336. https://doi.org/10.1207/S15327957PSPR0704_05

de Beaumont, G., & deTocqueville, A. (1833). *On the penitentiary system in the United States and its application in France*, ed. and trans. Francis Lieber. Philadelphia, PA: Carey, Lea and Blanchard.

De Castella, K., Platow, M. J., Wenzel, M., Okimoto, T. G., & Feather, N. T. (2011). Retribution or restoration? Anglo–Australian's views towards domestic violence involving Muslim and Anglo-Australian victims and offenders. *Psychology, Crime & Law*, *17*, 403–20. https://doi.org/10.1080/10683160903292253

Duckitt, J. (2006). Differential effects of right-wing authoritarianism and social dominance orientation on out-group attitudes and their mediation by threat from competitiveness to out-groups. *Personality and Social Psychology Bulletin*, *32*, 684–696. https://doi.org/10.1177/0146167205284282

Ellemers, N., Pagliaro, S., & Barreto, M. (2014). Morality and behavioural regulation in groups: A social identity approach. *European Review of Social Psychology*, *24*, 160–193. https://doi.org/10.1080/10463283.2013.841490

Ellemers, N., & Van den Bos, K. (2012). Morality in groups: On the social-regulatory functions of right and wrong. *Social and Personality Psychology Compass*, *6*, 878–889. https://doi.org/10.1111/spc3.12001

Everett, J. A. C., Faber, N., Savulescu, J., & Crockett, M. J. (2018). The costs of being consequentialist: Social inference from instrumental harm and impartial beneficence. *Journal of Experimental Social Psychology*, *79*, 200–216. https://doi.org/10.1016/j.jesp.2018.07.004

Fousiani, K., & Demoulin, S. (in press). The effects of group membership and perceived humanness of victims on motives for punishment and justice decisions. *Hellenic Journal of Psychology*.

Goldberg, J. H., Lerner, J. S., & Tetlock, P. E. (1999). Rage and reason: The psychology of the intuitive prosecutor. *European Journal of Social Psychology*, *29*, 781–795. https://doi.org/10.1002/(SICI)1099-0992(199908/09)29:5/6<781:AID-EJSP960>3.0.CO;2-3

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*, 20–26. https://doi.org/10.1027/1864-9335/a000004

Gollwitzer, M., & Van Prooijen, J.-W. (2016). Psychology of justice. In C. Sabbagh & M. Schmitt (Eds.), *Handbook of social justice theory and research* (pp. 61–82). New York, NY: Springer. https://doi.org/10.1007/978-1-4939-3216-0

Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review*, *108*, 814–834. https://doi.org/l0.l037//n033-295X.108.4.814

Hamilton, D. L., Sherman, S. J., & Rodgers, J. S. (2004). Perceiving the groupness of groups: Entitativity, homogeneity, essentialism, and stereotypes. In V. Yzerbyt, C. M. Judd & O. Corneille (Eds.), *The psychology of group perception: Perceived variability, entitativity, and essentialism* (pp. 39–60). New York, NY: Psychology Press.

Haslam, N. (2006). Dehumanization: an integrative review. *Personality and Social Psychology Review*, *10*, 252–264. https://doi.org/10.1207/s15327957pspr1003_4

Haslam, N. (2015). Dehumanization and intergroup relations. In M. Mikulincer, P. R. Shaver, J. F. Dovidio, J. A. Simpson, M. Mikulincer, P. R. Shaver, J. F. Dovidio & J. A. Simpson (Eds.), *APA handbook of personality and social psychology. Volume 2: Group processes* (pp. 295–314). Washington, DC: American Psychological Association. https://doi.org/10.1037/14342-000

Hayes, A. F. (2013). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach*. New York, NY: Guilford Press.

Hornsey, M. J., de Bruijn, P. M. H., Creed, J., Allen, J., Ariyanto, A., & Svensson, A. (2005). Keeping in in-house: How audience affects responses to group criticism. *European Journal of Social Psychology*, *35*, 291–312. https://doi.org/10.1002/ejsp.246

Hutchison, P., & Abrams, D. (2003). Ingroup identification moderates stereotype change in reaction to ingroup deviance. *European Journal of Social Psychology*, *33*, 497–506. https://doi.org/10.1002/(ISSN)1099-0992

Kahane, G., Everett, J. A. C., Earp, B. D., Caviola, L., Faber, N. S., Crockett, M. J., & Savulescu, J. (2018). Beyond sacrificial harm: A two-dimensional model of utilitarian psychology. *Psychological Review*, *125*, 131–164. https://doi.org/10.1037/rev0000093

Kant, I. (1797). The metaphysics of morals. In H. Reiss (Ed.), Translated by H. B. Nisbet. *Kant: Political writings* (2nd enl. ed.). Cambridge, UK: Cambridge University Press, 1991.

Kugler, M. B., Funk, F., Braun, J., Gollwitzer, M., Kay, A., & Darley, J. M. (2013). Differences in punitiveness across three cultures: A test of American exceptionalism in justice attitudes. *Journal of Criminal Law and Criminology*, *103*, 1071–1114.

Lacey, N., & Pickard, H. (2015). To blame or to forgive? Reconciling punishment and forgiveness in criminal justice *Oxford Journal of Legal Studies*, *35*, 665–698. https://doi.org/10.1093/ojls/gqv012

Leach, C. W., Ellemers, N., & Barreto, M. (2007). Group virtue: The importance of morality (vs. competence and sociability) in the positive evaluation of in-groups. *Journal of Personality and Social Psychology*, *93*, 234–249. https://doi.org/10.1037/0022-3514.93.2.234

Leach, C., van Zomeren, M., Zebel, S., Vliek, M., Pennekamp, S., Doosje, B., . . . Spears, R. (2008). Group-level self-definition and self-investment: A hierarchical (multicomponent) model of in-group identification. *Journal of Personality and Social Psychology*, *95*, 144–165. https://doi.org/10.1037/0022-3514.95.1.144

Lee, Y.-T., & Ottati, V. (1995). Perceived in-group homogeneity as a function of group membership salience and stereotype threat. *Personality and Social Psychology Bulletin*, *21*, 610–619. https://doi.org/10.1177/0146167295216007

Levine, J. M., & Moreland, R. L. (1994). Group socialization: Theory and research. In W. Stroebe & M. Hewstone (Eds.), *European review of social psychology* (Vol. 5, pp. 305–336). Chichester, UK: Wiley.

Levine, J. M., & Moreland, R. L. (2002). Group reactions to loyalty and disloyalty. In S. Thye & E. Lawler (Eds.), *Group cohesion, trust, and solidarity* (pp. 203–228). New York, NY: Elsevier Science.

Lewis, A., & Sherman, S. J. (2010). Perceived entitativity and the black-sheep effect: When will we denigrate negative ingroup members? *The Journal of Social Psychology*, *150*, 211–25. https://doi.org/10.1080/00224540903366388

Luhtanen, R., & Crocker, J. (1992). A collective self esteem scale Self evaluation of one's social identity. *Personality and social Psychology Bulletin*, *18*, 302–318. https://doi.org/10.1177/0146167292183006

Marques, J. M., & Paez, D. (1994). The "black sheep effect": Social categorization, rejection of ingroup deviates, and perception of group variability. In W. Stroebe & M. Hewstone (Eds.), *European review of social psychology* (Vol. *5*, pp. 37–68). New York, NY: John Wiley.

Marques, J. M., & Yzerbyt, V. Y. (1988). The black sheep effect: Judgmental extremity towards ingroup members in inter- and intragroup situations. *European Journal of Social Psychology*, *18*, 287–292. https://doi.org/10.1002/ejsp.2420180308

Marques, J. M., Yzerbyt, V. Y., & Leyens, J-Ph (1988). The black sheep effect: Judgmental extremity towards ingroup members as a function of group identification. *European Journal of Social Psychology*, *18*, 1–16. https://doi.org/10.1002/ejsp.2420180308

Meeussen, L., Phalet, K., Meeus, J., Van Acker, K., Montreuil, A., & Bourhis, R. (2012). "They are all the same": Low perceived typicality and outgroup disapproval as buffers of intergroup threat in mass media. *International Journal of Intercultural Relations*, *37*, 146–158. https://doi.org/10.1016/j.ijintrel.2012.05.002

Muller, D., Judd, C. M., & Yzerbyt, V. Y. (2005). When moderation is mediated and mediation is moderated. *Journal of Personality and Social Psychology*, *89*(6), 852–863. https://doi.org/10.1037/0022-3514.89.6.852

Nagin, D. S. (1998). Criminal deterrence research at the outset of the twenty-first century. In M. Tonry (Ed.), *Crime and justice: A review of research* (Vol. *23*). Chicago, IL: University of Chicago Press.

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

Pagliaro, S., Brambilla, M., Sacchi, S., D'Angelo, M., & Ellemers, N. (2013). Initial impressions determine behaviours: Morality predicts the willingness to help newcomers. *Journal of Business Ethics*, *117*, 37–44. https://doi.org/10.1007/s10551-012-1508-y

Pettigrew, T. F., Wagner, U., & Christ, O. (2010). Population ratios and prejudice: Modelling both contact and threat effects. *Journal of Ethnic and Migration Studies*, *36*, 635–650. https://doi.org/10.1080/13691830903516034

Piff, P. K., Martinez, A. G., & Keltner, D. (2012). Me against we: In-group transgression, collective shame, and in-group-directed hostility. *Cognition and Emotion*, *26*, 634–649. https://doi.org/10.1080/02699931.2011.595394

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

Rullo, M., Presaghi, F., & Livi, S. (2015). Reactions to ingroup and outgroup. Deviants: An experimental group paradigm for black sheep effect. *PLoS ONE*, *10*(5), e0125605. https://doi.org/10.1371/journal.pone.0125605

Saleilles, R. (1898). *L'individualisation de la peine*. Paris, France: Etude de criminalité sociale.

Sankaran, S., Sekerdej, M., & Von Hecker, U. (2017). The role of Indian caste identity and caste inconsistent norms on status representation. *Frontiers in Psychology: Personality and Social Psychology*, *8*, 1–14. https://doi.org/10.3389/fpsyg.2017.00487

Stephan, W. G., Boniecki, K. A., Ybarra, O., Bettencourt, A., Ervin, K. S., Jackson, L. A., McNatt, P. S., & Renfro, C. L. (2002). The role of threats in the racial attitudes of Blacks and Whites. *Personality and Social Psychology Bulletin*, *28*, 1242–1254. https://doi.org/10.1177/01461672022812009

Van der Lee, R. A., Ellemers, N., Scheepers, D. T., & Rutjens, B. (2017). In or out? How the morality (vs. competence) of prospective group members affects acceptance and rejection. *European Journal of Social Psychology*, *47*, 748–762. https://doi.org/10.1002/ejsp.2269

Van der Toorn, J., Ellemers, N., & Doosje, B. (2015). The threat of moral transgression: The impact of group membership and moral opportunity. *European Journal of Social Psychology*, *45*, 609–622. https://doi.org/10.1002/ejsp.2119

Van Leeuwen, E., van den Bosch, M., Castano, E., & Hopman, P. (2010). Dealing with deviants: The effectiveness of rejection, denial, and apologies on protecting the public image of a group. *European Journal of Social Psychology*, *40*, 282–299. https://doi.org/10.1002/ejsp.622

Van Prooijen, J.-W. (2018). *The moral punishment instinct*. New York, NY: Oxford University Press.

Van Prooijen, J.-W., & Lam, J. (2007). Retributive justice and social categorizations: The perceived fairness of punishment depends on intergroup status. *European Journal of Social Psychology*, *37*, 1244–1255. https://doi.org/10.1002/ejsp.421

Verkuyten, M. (2005). Ethnic group identification and group evaluation among minority and majority groups: Testing the multiculturalism hypothesis. *Journal of Personality and Social Psychology*, *88*(1), 121–138. https://doi.org/10.1037/0022-3514.88.1.121

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*, 450–470. https://doi.org/10.1007/s11211-006-0028-2

Yzerbyt, V. Y., Castano, E., Leyens, J-Ph, & Paladino, P. (2000). The primacy of the ingroup: The interplay of entitativity and identification. In W. Stroebe & M. Hewstone (Eds.), *European review of social psychology* (Vol. *11*, pp. 257–295). Chichester, UK: Wiley.

Yzerbyt, V. Y., & Demoulin, S. (2010). Intergroup relations. In S. T. Fiske, D. T. Gilbert & G. Lindzey (Eds.), *The handbook of social psychology* (5th ed.), Vol. *2* (pp. 1024–1083). Hoboken, NJ: Wiley.

Yzerbyt, V., Muller, D., Batailler, C., & Judd, C. M. (2018). New recommendations for testing indirect effects in mediational models: The need to report and test component paths. *Journal of Personality and Social Psychology*, *115*, 929–943. https://doi.org/10.1037/pspa0000132

Zehr, H. (1997). Restorative justice: The concept. *Corrections Today*, *59*, 68–70.

Received 31 October 2018; revised version received 18 December 2018

Supporting Information

The following supporting information may be found in the online edition of the article:

Appendix S1. Online Supplemental Material - Study vignettes and measures..