Judgment and Decision Making, Vol. 12, No. 4, July 2017, pp. 396–407

When is it appropriate to reprimand a norm violation? The roles of anger, behavioral consequences, violation severity, and social distance

Kimmo Eriksson*† Per A. Andersson‡ Pontus Strimling§†

Abstract

Experiments on economic games typically fail to find positive reputational effects of using peer punishment of selfish behavior in social dilemmas. Theorists had expected positive reputational effects because of the potentially beneficial consequences that punishment may have on norm violators' behavior. Going beyond the game-theoretic paradigm, we used vignettes to study how various social factors influence approval ratings of a peer who reprimands a violator of a group-beneficial norm. We found that ratings declined when punishers showed anger, and this effect was mediated by perceived aggressiveness. Thus the same emotions that motivate peer punishers may make them come across as aggressive, to the detriment of their reputation. However, the negative effect of showing anger disappeared when the norm violation was sufficiently severe. Ratings of punishers were also influenced by social distance, such that it is less appropriate for a stranger than a friend to reprimand a violator. In sum, peer punisher ratings were very high for a friend reprimanding a severe norm violation, but particularly poor for a stranger showing anger at a mild norm violation. We found no effect on ratings of whether the reprimand had the beneficial consequence of changing the violator's behavior. Our findings provide insight into how peer punishers can avoid negative reputational effects. They also point to the importance of going beyond economic games when studying peer punishment.

Keywords: peer punishment, social distance, consequentialism, aggression, anger.

1 Introduction

Someone plays a loud movie in a crowded train. Someone litters in a public area. Someone jumps the line to a music club. These are examples of social norm violations with negative consequences for other people in the same group or same environment. Such behaviors have been termed "uncivil" (e.g., Brauer & Chaurand, 2010). In a series of field studies, Chaurand and Brauer (2008) investigated determinants of whether a bystander speaks up against uncivil behavior. They identified three factors in bystanders and their relation to the situation and the norm violator: speaking up is more likely if the bystander is angry, if the bystander feels responsible for speaking up, and if the bystander perceives he or she has the legitimacy to do it. These factors explained variation between individuals as well as between situations.[1]

The importance of responsibility and legitimacy indicate that speaking up against a norm violation is in itself subject to social norms. Such norms have been referred to as norms about punishment of norm violations (Strimling & Eriksson, 2014), rules of politeness (Felson, 1981), second-order norms (Elster, 1989), or simply meta-norms (Axelrod, 1986). A field experiment of such meta-norms was recently conducted at a train station in Germany, where an actor playing the role of "violator" dropped a coffee cup on the platform and another actor playing the role of "punisher" told the violator to pick up his garbage (Balafoutas, Nikiforakis & Rockenbach, 2014). When the punisher then, seemingly by accident, dropped some books in front a bystander, he did not receive help more frequently than in a control condition without punishment. Thus, speaking up was not socially rewarded in this setting. This finding in the field is consistent with the results from a line of research that has studied how people react to "peer punishers", a term for individuals who without formal authority respond negatively to an uncivil behavior. Peer punishers have typically not been judged more positively than non-punishers, either in economic games (e.g., Cinyabuguma et al., 2006; Eriksson et al., 2017; Kiyonari & Barclay, 2008), in vignettes (Strimling & Eriksson, 2014), or in computer animations (Eriksson, Andersson & Strimling, 2016). However, research on judgments of peer punishers has typically neglected the role of various social factors that may influence the judgment. The aim of the present paper is to examine the roles of several such factors.

This research was funded by the Swedish Research Council [grant number 2009–2390] and the Knut and Alice Wallenberg Foundation [grant number 2015.0005].

*School of Education, Culture and Communication, Mälardalen University, Västerås, Sweden. Email: kimmo.eriksson@mdh.se.

†Centre for the Study of Cultural Evolution, Stockholm University, Stockholm, Sweden.

‡Division of Economics, Department of Management and Engineering, Linköping University, Linköping, Sweden.

§Institute for Futures Studies, Stockholm, Sweden.

[1] Sabini and Silver (1978) report a similar analysis.

1.1 Showing anger

Our main focus will be the role of showing anger. We know from previous research that people who get angry are more likely to speak up against a norm violation (Chaurand & Brauer, 2008). But is it appropriate for them to show they are angry when they speak up? Or might that make them come across as aggressive?

The concepts of anger and aggression and the connection between them have been the subject of much work in psychology (e.g., Averill, 1983; Berkowitz, 1990; for a recent review, see Averill, 2012). Anger refers to an emotional state whereas aggression refers to a behavior that is harmful or threatening to others. Aggression that is driven by anger has been termed reactive aggression, to distinguish it from proactive aggression that is cold-blooded and calculated (Lochman et al., 2010). In studies of children, these two types of aggression have been found to have distinct social consequences: reactive aggression is associated with peer rejection, but proactive aggression is not (Dodge et al. 1997). Thus, it would seem that aggression is socially condemned mainly when it is accompanied with anger. Indeed, it seems plausible that the exact same behavior may come across as more aggressive, and therefore be more condemned, when there are signs of the actor being angry. Experimental research using manipulated facial expressions has found that signs of someone being angry make others judge that person as higher on dominance and lower on affiliation (Knutson, 1996), and judge the situation as more competitive and less cooperative (Van Doorn et al., 2012).

Consistent with these findings we predict that when peer punishers show anger as they reprimand a norm violation, they will be judged as more aggressive than when they do not show anger. Moreover, as reactive aggression tends to be socially condemned, we expect the reprimand to be rated as less appropriate when punishers are perceived as more aggressive. However, how inappropriate it is to show anger may be moderated by the severity of the norm violation. We turn to this factor next.

1.2 Severity of the norm violation

A follow-up to the above-mentioned field experiment of littering on a train platform studied how much bystanders punished a mild instance of violation of the norm against littering, in the form of a dropped coffee cup, compared to a more severe instance in the form of a dropped paper bag full of trash (Balafoutas, Nikiforakis & Rockenbach, 2016). A survey among passengers showed that they tended to believe that the more severe violation should be more strongly reprimanded. It seems plausible that a show of anger is perceived as a stronger reprimand. Thus, when the norm violation is more severe it may be generally more appropriate to reprimand it, and also less inappropriate to show anger when doing so.

In the actual experiment of Balafoutas et al. (2016), there was no difference in punishment rates between the two norm violations. A previous study similarly found no relationship between degree of deviance and social control (Brauer & Chekroun, 2005). Balafoutas et al. found the null effect to be explained by a tendency to perceive a greater risk of retaliation from the more severe norm violator. Note that our interest in the present paper lies in the general public's — not the violator's — appraisal of the peer punisher. The sentiments of the norm violator are surely not representative of the general public.

1.3 Social distance

In studies using economic games, researchers distinguish between "stranger treatments", in which each participant interact with a new set of players in each round, and "partner treatments", in which the same set of players repeatedly interact (Fehr & Gächter, 2000). Outside the laboratory, a corresponding distinction is that between strangers and friends, that is, the dimension of social distance. Returning to the framework of Chaurand and Brauer (2008), social distance may have bearing both on responsibility and legitimacy.

In situations in which there is variation in social distance, vignette studies indicate that it is more appropriate to speak up against a norm violation for those who are socially closer to the violator (Strimling & Eriksson, 2014). For instance, if there is both a friend and a stranger present, people tend to think that it is the friend who should speak up rather than the stranger. This could be due to a general effect of social distance, such that it is always more appropriate for a friend than for a stranger to reprimand a norm violator. A complementary explanation is that social distance works as a coordination device, to say who should punish in a given situation (Eriksson, Strimling & Ehn, 2013).

Social distance may also interact with anger. Namely, if peer punishment of norm violations is generally more appropriate between friends than between strangers, it may also be less inappropriate to show anger when reprimanding a friend than when reprimanding a stranger.

1.4 Beneficial consequences

The literature on peer punishment of uncivil behavior is dominated by research on social dilemmas. These are situations in which there is a selfish motive for every individual to behave in a way that is bad for the group. Many uncivil behaviors can be shoe-horned into the social dilemma model: it is conceivable that everyone's genuine preference would be to be able to watch a loud movie on a train, or to just drop their coffee cup on the platform instead of making the effort of finding a trash can, etc. In the laboratory, economic games are used to create more unambiguous social dilemmas. In such games, group-beneficial behavior may

sometimes be upheld by means of peer punishment (Fehr & Gächter, 2000; for a meta-analysis, see Balliet, Mulder & Van Lange, 2011). Because peer punishment potentially has beneficial consequences for the group, many theorists have regarded the provision of punishment as a public good (e.g., Gardner & West, 2004; Henrich & Boyd, 2001; Nakao & Machery, 2012). However, the peer punishers themselves do not seem to care much about the consequences for the group (Eriksson, et al., 2014). It has not been investigated how much lay people care about these beneficial consequences when they make social judgments of peer punishers.

If instead we turn to the literature on legal punishment, related questions have been studied. Sunstein, Schkade and Kahneman (2000) pointed out that while optimal deterrence is the goal in economic theory of punishment, it may not be an important criterion in ordinary people's judgments of punishment. Indeed, in two studies they found that people tend to not take deterrence effects into account when proposing punishments, and tend to reject optimally deterrent punishments. Similar findings of limited attention to deterrence effects have been obtained in several other studies (e.g., Baron & Ritov, 1993, 2009; Sunstein, Kahneman & Schkade, 1998). Nonconsequentialist tendencies in judgments of legal punishments may be part of a more general phenomenon (Baron, 1994). In particular, such tendencies should carry over to informal contexts. We therefore do not expect judgments of peer punishment to be strongly influenced by whether the beneficial consequence of changing the norm violator's behavior in fact was realized. This prediction stands in contrast to the social dilemma perspective, according to which the *raison d'être* of peer punishment is its beneficial consequences.

1.5 Outline of studies

Four studies are reported in this paper, all using the methodology of vignettes. Each vignette presents a situation in which a punisher reprimands a norm-breaker for behaving selfishly. Vignettes were manipulated across conditions that differed between studies. Studies 1, 3 and 4 used a US sample of MTurk users. Study 2 replicated Study 1 with a sample of Swedish students.

Studies 1 and 2 concern two questions about judgments of the appropriateness of peer punishment of rather mild norm violations: (1) Are they influenced by whether the punisher shows anger, and is this influence mediated by perceived aggression? (2) Are they influenced by whether the punishment has beneficial consequences on the norm-violator's behavior?

Studies 3 and 4 use a greater range of norm violations to focus on the effect of anger and how it may be moderated by the severity of the norm violation. Study 4 also manipulates social distance (friends vs. strangers).

2 Study 1

The aim of the first study was to examine how the appropriateness of a reprimand depends on whether the punisher shows anger and whether there are beneficial consequences in the form of a change of the norm violator's behavior.

2.1 Method

Participants. Participants were 400 adults (58% male, age ranging from 19 to 75 years with a mean of 36 years) recruited among American users of Amazon Mechanical Turk at a fee of 0.50 US dollar.

Materials and procedure. Participants were directed at random to one of four versions of an online form, yielding approximately 100 participants per condition of a two-by-two between-participant design. The form presented three scenarios in which someone (the violator) behaved uncivilly: by littering in a park, by taking a too large piece of cake, or by watching a loud movie on a crowded train. In each scenario someone else (the punisher) reprimanded the violator for bad behavior. Details of the scenarios varied according to the two-by-two design: [Anger, No anger] × [Beneficial, Not beneficial]. In the Anger + Beneficial condition the scenarios read as follows:

> [Litter] Adrian is in the park waiting to meet his friend Burt. While waiting he is eating. Burt arrives just as Adrian is about to throw some packaging on the ground. Burt realizes that Adrian was about to litter. Burt gets angry. Burt reprimands Adrian, telling him that littering is bad behavior. Adrian then refrains from littering.
>
> [Cake] Clara is at a dinner party and a cake is served. There is enough cake for a small piece to each person. Clara is about to take a much larger piece than that when Diana, who has been to the bathroom, comes for cake. Diana realizes that Clara was about to take more than her share. Diana gets angry. Diana reprimands Clara, telling her that taking more than everyone else gets is bad behavior. Clara then refrains from taking more than a small piece.
>
> [Movie] Edwin is traveling home on a crowded train. As he gets on the train he starts watching a loud movie on his phone without using earphones. At that moment he meets Fay, a passenger who is leaving the train. Fay realizes that Edwin is about to watch a loud movie on a crowded train. Fay gets angry. Fay reprimands Edwin, telling him that this is bad behavior. Edwin then refrains from watching the loud movie.

The No anger condition was obtained by removing the sentences "Burt/Diana/Fay gets angry." The Not beneficial condition was obtained by changing the final sen-

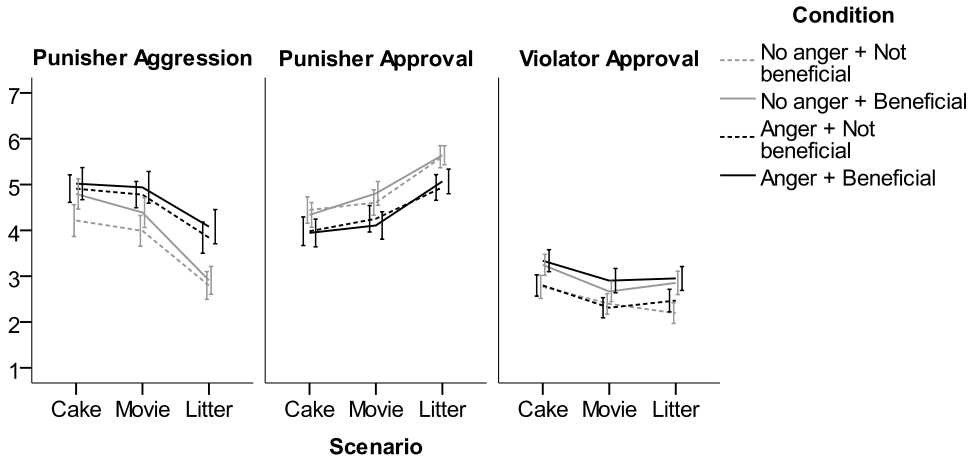

FIGURE 1: Mean ratings of punisher aggression, punisher approval, and violator approval, depending on scenario and condition in Study 1. All ratings used a scale from 1 to 7. Error bars indicate 95% confidence intervals.

tences to "Adrian/Clara/Edwin decides to litter/take the large piece/watch the loud movie anyway."

Each scenario was followed by ratings of the violator and the punisher. Both were rated on a three-item approval scale from Eriksson et al. (2016):

1. I think [...]'s behavior was appropriate.
2. I would like to spend time with a person who behaves like [...].
3. (*reverse coded*) If a person who behaves like [...] belonged to my group I would consider that person to be a problem (rather than an asset) for the group.

The scale showed adequate internal consistency (Cronbach's alpha > .77 in every case).

The punisher, but not the violator, was also rated on a fourth item: "I saw [...]'s behavior as aggressive." Ratings were done on a seven-point response scale from 1 = *strongly disagree* to 7 = *strongly agree*.

2.2 Analysis and results

For each combination of conditions in each scenario, Figure 1 shows mean ratings and 95% confidence intervals for punisher aggression, punisher approval, and violator approval.

Punisher approval. We analyzed punisher approval using a mixed-design ANOVA with scenario (Cake, Movie, Litter) as a within-subjects factor and with anger (Anger, No anger) and consequences (Beneficial, Not beneficial) as between-subjects factors. There was a large main effect of scenario, $F(2, 792) = 129.78$, $p < .001$, $\eta_p^2 = .25$. Figure 1 shows that punisher approval was higher in the Litter scenario than in the other two scenarios.

Our main focus here is the effect of anger. There was a medium-sized main effect of anger, $F(1, 396) = 22.71$, $p < .001$, $\eta_p^2 = .05$. Figure 1 shows that punisher approval was lower in the Anger condition than in the No anger condition.

There was no significant main effect of consequences, $F(1, 396) = 0.01$, $p = .92$, $\eta_p^2 = .000$, and no significant interactions between factors, all $F \leq 1.41$, $p \geq .24$, $\eta_p^2 \leq .004$.

Mediation analysis. Figure 1 shows that punisher aggression ratings were higher in the Anger condition in every scenario. To test the prediction that perceived punisher aggression mediates the relationship between anger condition and approval, we used the basic mediation model of the PROCESS macro in SPSS (Hayes, 2013) with 5000 bootstrapped samples. This macro calculates a series of regression coefficients representing the path from the independent variable X to the mediator M (denoted by a), the path (the coefficient with X as a covariate) from M to the dependent variable Y (denoted by b), the direct effect of X on Y (denoted by c'), the total effect of X on Y (denoted by c), a bootstrapped bias-corrected 95% confidence interval of the indirect effect of X on Y through M (the product ab), and the ratio of the indirect effect to the total effect (P_M). Table 1 reports these estimates for every scenario. In each scenario the indirect effect through the mediator was significant and accounted for most of the total effect.

Violator approval. Although violator approval is not the focus of this study, it is worth noting that it was not influenced by conditions in the same way as punisher approval. Figure 1 shows that when the reprimand had beneficial consequences on the violator's behavior, violator approval increased. In contrast, whether the punisher showed anger had no effect on violator approval.

TABLE 1: Results of mediation analysis of the effect of anger on approval via perceived aggression in Study 1.

Scenario	a	b	c'	c	ab [BCa CI]	P_M
Cake	0.47^{**}	-0.56^{***}	-0.17^{ns}	-0.43^{**}	−0.26 [−0.48, −0.08]	.61
Movie	0.68^{***}	-0.50^{***}	-0.19^{ns}	-0.52^{***}	−0.34 [−0.52, −0.17]	.64
Litter	1.10^{***}	-0.41^{***}	-0.17^{ns}	-0.62^{***}	−0.46 [−0.63, −0.30]	.73

^{ns} $p > .1$, ** $p < .01$, *** $p < .001$

Note: For the basic mediation model (X to Y through M), a and b are regression coefficients for the paths from X to M and from M to Y, respectively, such that ab is the indirect effect of X on Y through M. The direct effect of X on Y is c' and the total effect is c. The ratio of the indirect effect to the total effect is P_M.

2.3 Discussion

This study found support for the hypothesis that, even if the reprimand is the same, an angry punisher is perceived as more aggressive and therefore judged as behaving less appropriately. The same effect of anger was found across three scenarios that differed somewhat in baseline punisher approval.

In contrast to the clear effect of the punisher's anger, we found no effect on punisher approval from the consequences of the reprimand. Whereas judgments of norm violators were less negative when they changed their behavior, judgments of punishers were not affected.

Our first concern was whether these findings depended on the study being run on MTurk, where users may gain experience from participating in a large number of studies (Dance, 2015) and where data reliability may be lower than in other samples (Rouse, 2015). We address this limitation in Study 2.

3 Study 2

The aim of the second study was to replicate Study 1 in another sample.

3.1 Method

Participants. Participants were 203 adults (27% male, age ranging from 18 to 59 years with a mean of 27 years) recruited from a pool of students at Swedish universities who had previously signed up as willing to participate in online studies at the website vetenskaponline.se. Participation was rewarded by 100 SEK (roughly 11 USD).

Materials and procedure. Participants filled out an online form in several parts. The first part, involving rating of animations, is not analyzed here; it was a study for an unrelated cross-cultural project on peer punishment and did not include an experimental manipulation. The second part, which we analyze here, was a Swedish translation of Study 1. (Translation available on request.)

3.2 Analysis and results

We replicated the analysis performed in Study 1. Figure 2 shows mean ratings and 95% confidence intervals for punisher aggression, punisher approval, and violator approval in each scenario.

Punisher approval. The same mixed-design ANOVA of punisher approval replicated the main effect of scenario, $F(2, 398) = 39.03$, $p < .001$, $\eta_p^2 = .16$, with approval highest in the Litter scenario, and the main effect of anger, $F(1, 199) = 11.76$, $p < .001$, $\eta_p^2 = .06$, with approval lower in the Anger condition. As in Study 1, there was no significant main effect of consequences, $F(1, 199) = 0.25$, $p = .62$, $\eta_p^2 = .000$, and no significant interactions between factors, all $F \leq 1.21$, $p \geq .30$, $\eta_p^2 \leq .006$.

Mediation analysis. As in Study 1, punisher aggression ratings were higher in the Anger condition in every scenario. The same mediation analysis as in Study 1 yielded similar, but consistently stronger, results. See Table 2. In every scenario the indirect effect of anger condition on punisher approval through perceived punisher aggression was significant and accounted for the entire total effect.

Violator approval. Figure 2 shows that also violator approval exibited the same pattern as in Study 1: violator approval increased when the violator changed behavior after the reprimand, but was unaffected by punisher anger.

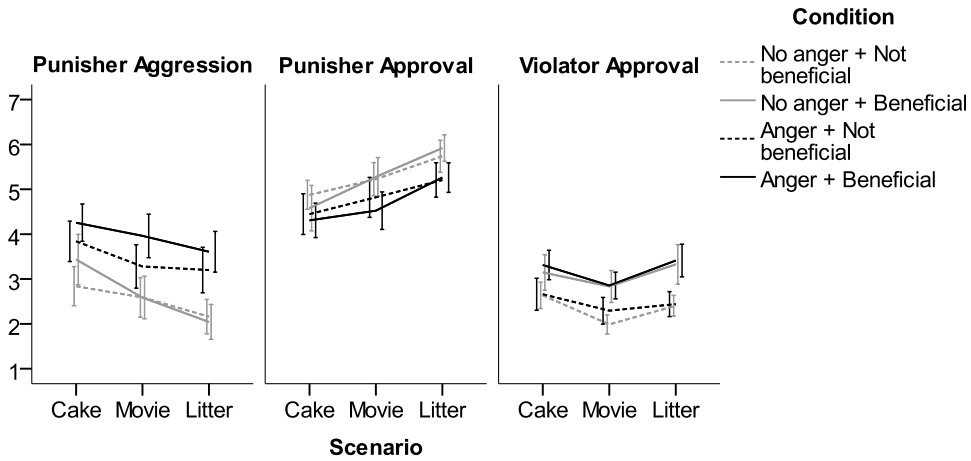

FIGURE 2: Mean ratings of punisher aggression, punisher approval, and violator approval, depending on scenario and condition in Study 2. All ratings used a scale from 1 to 7. Error bars indicate 95% confidence intervals.

TABLE 2: Results of mediation analysis of the effect of anger on approval via perceived aggression in Study 2.

Scenario	a	b	c'	c	ab [BCa CI]	P_M
Cake	0.94^{***}	-0.57^{***}	0.17^{ns}	$-0.37^{\dagger}$	-0.54 [$-0.82, -0.28$]	1.48
Movie	1.04^{***}	-0.56^{***}	0.01^{ns}	-0.58^{**}	-0.58 [$-0.89, -0.31$]	1.01
Litter	1.30^{***}	-0.51^{***}	0.08^{ns}	-0.59^{***}	-0.66 [$-0.96, -0.42$]	1.13

^{ns} $p > .1$, † $p < .1$, ** $p < .01$, *** $p < .001$

Note: Symbols are explained in Table 1.

3.3 Discussion

Study 2 replicated the findings of Study 1 in a non-MTurk sample. This increases our confidence in the effect of the punisher's anger on judgments of the appropriateness of the punisher's behavior, via perceived aggressiveness. However, none of the scenarios involved a really severe norm violation. As we argued in the introduction, it is possible that anger and aggression are more condoned when the norm violation is more severe. Scenarios alternated between friends and strangers, and between men and women, which may confound comparisons between scenarios.

Another limitation is that the anger manipulation was ambiguous. The No anger condition did not explicitly say that the punisher did not show anger, and the Anger condition did not explicitly say that the punisher showed anger.

A further limitation was that we did not obtain ratings of someone who did *not* reprimand in the same scenarios. This means that we do not know whether reprimanding in these scenarios is judged as more or less appropriate than not reprimanding at all.

Finally, the scenarios all described reprimands to prevent a norm violation. This contrasts with most of the literature on peer punishment, which has been focused on punishment *after* a norm violation.

4 Study 3

The aim of the third study was to address the above-mentioned limitations.

4.1 Method

Participants. Participants were 151 adults (54% male, age ranging from 19 to 70 years with a mean of 34 years) recruited among American users of Amazon Mechanical Turk at a fee of 0.50 US dollar.

Materials. To include a wider range of norm violations, Study 3 increased the number of scenarios to eight (Music, Cake, Meet, Skate, Litter, Line, Drive, Vandal). Every scenario existed in three different versions (Not punish, Not show anger, Show anger). To make scenarios otherwise comparable, we held gender and relation constant: every

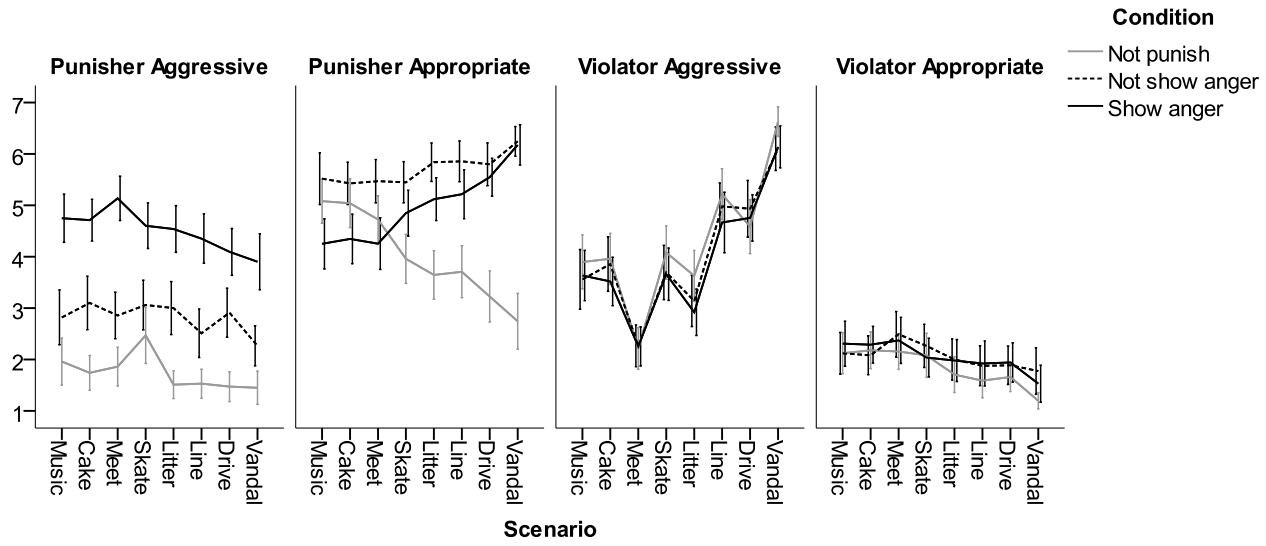

FIGURE 3: Mean ratings of punisher aggression, punisher approval, violator aggression, and violator approval, depending on scenario and condition in Study 3. All ratings used a scale from 1 to 7. Error bars indicate 95% confidence intervals.

scenario involved male friends. The basic scenarios read as follows.

> [Skate] Daniel is in a shopping mall, carrying his skateboard. Daniel then decides to ride his skateboard inside of a shop in the mall. As he exits the shop he picks up his skateboard again. His friend Nathan realizes that Daniel has been riding his skateboard inside of the shop. Nathan gets angry
>
> [Drive] Mike is driving his friend Nick into town on a 45 mph road. They both see a sign saying "School zone 20 mph". After they have passed the entire school zone Nick realizes that Mike did not slow down but kept going at 45 mph. Nick gets angry
>
> [Cake] Larry is at a dinner party and a cake is served. There is enough cake for a small piece to each person. Larry takes a much larger piece than that. His friend William realizes that Larry has taken more than his share and eaten it. William gets angry
>
> [Line] Isaac wants to get a good seat at music club. There are twenty people waiting in line, but Isaac walks straight to the front of the line. His friend Jack, who is standing in line, realizes that Isaac has jumped the line and gotten inside. Jack gets angry
>
> [Meet] George is in a team meeting but keeps fiddling with his mobile phone during the meeting. After the meeting his friend Henry realizes that George was not paying attention during the meeting. Henry gets angry
>
> [Music] Edwin is traveling home on a crowded train. When he gets on the train he watches a loud movie on his phone without using earphones. As he is getting off the train he meets Lucas, a friend who is entering the train. Lucas realizes that Edwin has watched a loud movie on a crowded train. Lucas gets angry
>
> [Litter] Adrian is in the park waiting to meet his friend Burt. While waiting he is eating. Burt arrives just as Adrian is finished eating and has thrown some packaging on the ground. Burt realizes that Adrian has been littering. Burt gets angry
>
> [Vandal] Karl has made up his mind to vandalize the town's Christmas tree by setting fire to it. His friend Lenny realizes that Karl has set fire to the tree and that the tree has burned down. Lenny gets angry

Conditions differed only in how the last sentence was completed:

> [Show anger] ...and, letting the anger show, reprimands <*the violator*>, telling him that it is bad behavior.
>
> [Not show anger] ...and, not letting the anger show, reprimands <*the violator*>, telling him that it is bad behavior.
>
> [Not punish] ...but does not let the anger show and does not reprimand <*the violator*>.

Procedure. Participants were directed to an online form that presented the eight scenarios in random order. Which version was presented to the participant was randomized for each scenario.

As the number of scenarios was much larger than in the previous studies, we simplified the approval rating by using only the first item (i.e., whether the behavior was appropriate). The aggressiveness rating preceded the appropriateness rating. Appropriateness and aggressiveness ratings were done both for the punisher and the violator.

TABLE 3: Results of mediation analysis of the effect of anger on approval via perceived aggression in Study 3.

Scenario	a	b	c'	c	ab [BCa CI]	P_M
Music	1.93^{***}	-0.45^{***}	-0.40^{ns}	-1.27^{***}	−0.87 [−1.46, −0.39]	0.68
Cake	1.61^{***}	-0.53^{***}	-0.23^{ns}	-1.08^{***}	−0.85 [−1.39, −0.48]	0.79
Meet	2.28^{***}	-0.39^{***}	-0.33^{ns}	-1.21^{***}	−0.89 [−1.54, −0.40]	0.73
Skate	1.54^{***}	-0.39^{***}	-0.00^{ns}	-0.60^{*}	−0.60 [−1.11, −0.26]	1.00
Litter	1.54^{***}	-0.35^{***}	-0.18^{ns}	-0.72^{*}	−0.53 [−0.92, −0.29]	0.74
Line	1.84^{***}	-0.35^{***}	0.02^{ns}	-0.64^{*}	−0.66 [−1.24, −0.28]	1.02
Drive	1.18^{***}	-0.26^{**}	0.06^{ns}	-0.25^{ns}	−0.31 [−0.66, −0.09]	1.22
Vandal	1.64^{***}	-0.19^{*}	0.24^{ns}	-0.07^{ns}	−0.31 [−0.73, −0.08]	4.53

^{ns} $p > .1$, * $p < .05$, ** $p < .01$, *** $p < .001$

Note: Symbols are explained in Table 1.

4.2 Analysis and results

Figure 3 shows mean ratings and 95% confidence intervals for the four ratings of each scenario. Our main interest lies in ratings of the punisher. However, first note that violator appropriateness was very low for all scenarios (indicating that they were all considered to be norm violations), whereas there was a remarkable variation between scenarios in the perceived aggressiveness of the violator (indicating that some norm violations, such as fiddling with your phone in a meeting, are not seen as aggressive). Note also that ratings of the violator were not sensitive to whether the violation was punished.

The effect of showing anger on punishment appropriateness. Replicating the previous studies, a reprimand was typically rated as more aggressive, and less appropriate, when the punisher showed anger than when he did not. See Figure 3. The effect of showing anger on aggressiveness ratings was roughly constant across scenarios. The effect on appropriateness, however, essentially vanished in the Drive and Vandal scenarios. Note that these scenarios involved the most serious norm violations (speeding outside a school and vandalism).

Table 3 reports the results of mediation analyses, conducted as in the previous studies. The c column shows that the effect on appropriateness of showing anger was significant for all scenarios except Drive and Vandal. In every scenario, the indirect effect through perceived punisher aggression (ab) was significant and accounted for most or all of the total effect (ratio given by P_M).

Punishing vs. nonpunishing. We now move on to ratings of a nonpunisher. Figure 3 shows that nonpunishers are consistently rated as less aggressive than punishers, but they are not necessarily rated as behaving more appropriately. For mild norm violations, such as someone taking too much cake, it was equally appropriate to reprimand as to refrain from reprimanding. For more severe norm violations, however, nonpunishment was not seen as appropriate.

It is interesting to note that differences in the severity of norm violations seem to be measured better by ratings of the appropriateness of nonpunishment than by ratings of the appropriateness of the violation (which were very low across scenarios). To illustrate how the effect of showing anger decreased with the severity of the norm violation we therefore plotted the mean difference in appropriateness between not showing and showing anger against the mean appropriateness of nonpunishment, see Figure 4. The correlation was very high, $r = .96$, $p < .001$.

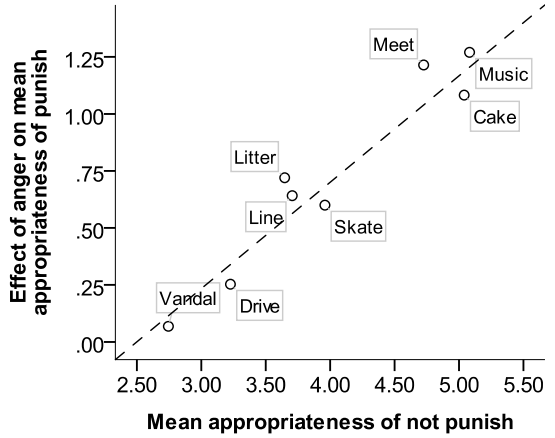

FIGURE 4: Dotplot of how the mean effect of showing anger on the appropriateness of punishment depends on the mean appropriateness of nonpunishment.

4.3 Discussion

The design of this study differed in several details to the previous studies (e.g., reprimands were given after the norm violations instead of attempting to prevent them, and punishers were explicitly described as showing anger or not showing anger). The main findings of the previous studies turned out to be robust to these details. Thus, for norm violations that are not too severe it is less appropriate to show anger than to not show anger, and this effect is mediated by perceived aggression.

This study also yielded new insights. When norm violations were more severe, such as vandalism or speeding outside a school, the negative effect of showing anger upon appropriateness ratings vanished. Moreover, to refrain from punishment was seen as appropriate only when the norm violation was mild.

Limitations of this study include that the punisher was a friend of the norm violator and that there was only one potential punisher. The literature on peer punishment is mainly concerned with punishment between strangers and with the case where there are several potential punishers.

5 Study 4

We conducted a fourth study to address the limitations identified above.

5.1 Method

Participants. Participants were 203 adults (54% male, age ranging from 18 to 73 years with a mean of 35 years) recruited among American users of Amazon Mechanical Turk at a fee of 0.50 US dollar.

Materials. Study 4 used the same basic eight scenarios as Study 3 (i.e., Music, Cake, Meet, Skate, Litter, Line, Drive, Vandal), but now included two bystanders: one person who decided not to punish and another one who did. Scenarios were manipulated in two ways. First, the punisher was described either as showing anger or as not showing anger. Second, the nonpunisher and the punisher were both described either as friends or as strangers of the violator. Here is an example of what the exact changes to the basic scenarios looked like:

> [Skate: Friends] ... His friends Nathan and Trevor realize that Daniel has been riding his skateboard inside of the shop. Trevor decides to let it go but Nathan gets angry ...
>
> [Skate: Strangers] ... Two strangers, Nathan and Trevor, realize that Daniel has been riding his skateboard inside of the shop. Trevor decides to let it go but Nathan gets angry ...

Procedure. Participants were directed to an online form that presented the eight scenarios in a different random order for each participant. Participants in one condition read all scenarios in the Friends version, participants in another condition read all scenarios in the Strangers version. Whether the punisher was presented as showing anger or not showing anger was randomized for each scenario.

As the number of actors to be rated was larger in this study, we dropped the aggressiveness ratings and focused on the appropriateness ratings. Participants rated how appropriately the nonpunisher, the punisher, and the violator had behaved. In case the punisher had been presented as showing anger, participants were also asked to imagine that the punisher would not have let his anger show, and give a separate rating for this case. In case the punisher had been presented as not showing anger, participants were instead asked to imagine that he had showed anger. Thus, for every participant we obtained four ratings: the violator, a nonpunisher, a punisher who did not show anger, and a punisher who did.

5.2 Analysis and results

We do not present the violator ratings; they were essentially the same as in Study 3 and indistinguishable between the Friends and Stranger conditions. Figure 5 shows mean ratings and 95% confidence intervals for the ratings of the nonpunisher and the two versions of the punisher.

Effects on punisher ratings of scenario, anger, and social distance. We ran a mixed-design ANOVA with scenario (Music, Cake, Meet, Skate, Litter, Line, Drive, Vandal) and anger (Show, Not show) as within-subjects factors, and with social distance (Friends, Strangers) as a between-subjects factor. There was a large main effect of scenario, $F(7, 1407) = 57.80$, $p < .001$, $\eta_p^2 = .22$, and a large main effect of anger, $F(1, 201) = 90.11$, $p < .001$, $\eta_p^2 = .31$, as well as a significant interaction between scenario and anger, $F(7, 1407) = 10.07$, $p < .001$, $\eta_p^2 = .05$. These results describe the same findings as in Study 3: punisher ratings are higher in scenarios involving more severe norm violations and lower for punishers who show anger, and the effect of anger is lower for more severe norm violations.

There was no significant main effect of social distance, $F(1, 201) = 0.74$, $p = .39$, $\eta_p^2 = .00$. However, there was a significant interaction between social distance and anger, $F(1, 201) = 4.44$, $p = .036$, $\eta_p^2 = .022$. In Figure 5 this can be seen as a larger effect of anger in the Strangers condition than in the Friends condition. In other words, to show anger was less appropriate for a stranger than for a friend.

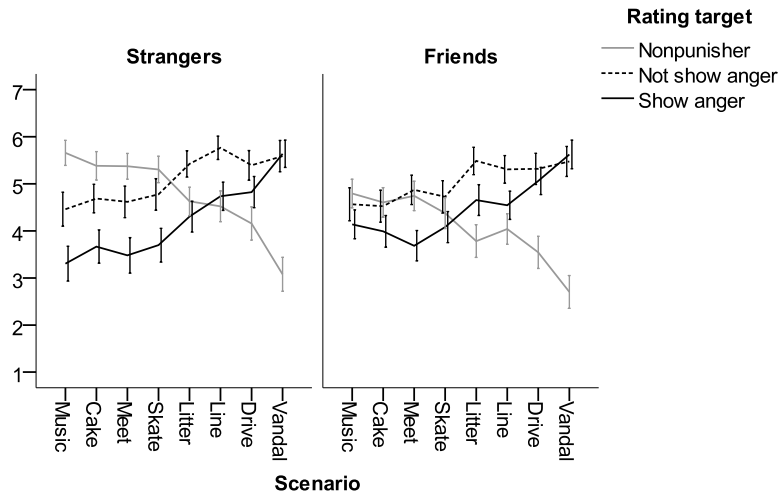

FIGURE 5: Mean ratings of the appropriateness of punishing with show of anger, punishing without showing anger, and not punishing at all, depending on scenario and condition in Study 4. All ratings used a scale from 1 to 7. Error bars indicate 95% confidence intervals.

Finally, there was a small but significant interaction between social distance and scenario, $F(7, 1407) = 2.46$, $p = .016$, $\eta_p^2 = .012$. In Figure 5 this can be seen as a somewhat larger effect of scenario in the Strangers condition than in the Friends condition. In other words, to punish a mild norm violation was somewhat less appropriate for a stranger than for a friend.

Effects of scenario and social distance on nonpunisher appropriateness. Next we turn to ratings of the appropriateness of *not* punishing a norm violation. We ran a mixed-design ANOVA with scenario as a within-subjects factor and social distance as a between-subjects factor. There was a large main effect of scenario, $F(7, 1407) = 80.01$, $p < .001$, $\eta_p^2 = .28$. As in Study 3, ratings of nonpunishers were much lower in scenarios involving more severe norm violations.

There was also medium-sized main effect of social distance, $F(1, 201) = 18.99$, $p < .001$, $\eta_p^2 = .09$. Figure 5 shows how nonpunishing was more appropriate for strangers than for friends. The effect of social distance was not significantly moderated by scenario, $F(7, 1407) = 1.27$, $p = .26$, $\eta_p^2 = .006$.

Punishers vs. nonpunishers In line with the findings reported so far, Figure 3 shows that strangers who did not punish mild norm violations were rated significantly higher than strangers who punished without showing anger. This was not observed in the Friends condition in this study, nor in Study 3.

Finally, as in Study 3 there was a very high correlation across scenarios between the mean effect of punishers showing anger and the mean appropriateness of nonpunishing (pooling the Friends and Strangers conditions), $r = .83$, $p = .012$. See Figure 6.

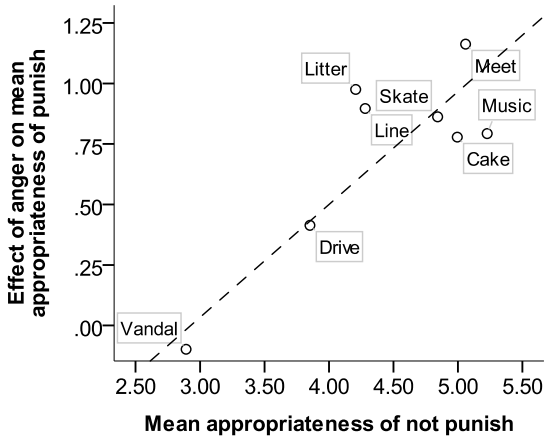

FIGURE 6: Dotplot of how the mean effect of showing anger on the appropriateness of punishment depends on the mean appropriateness of nonpunishment.

6 General discussion

In the introduction we discussed a growing body of research on norms about peer punishment in social dilemmas, and how it is shaped by a game theoretic research tradition. Against this backdrop, the present study aimed at enriching this research by examining the role of hostile emotions such as anger. It is known from field research that anger is an important determinant for bystanders speaking up against uncivil behavior (Chaurand & Brauer, 2008). In four studies we found that when peer punishers reacted to a norm vio-

lation with a verbal reprimand, they were judged differently depending on whether they showed anger. Specifically, approval ratings declined when punishers showed anger. Moreover, this effect was mediated by perceived aggressiveness. We conclude that the same emotions that motivate peer punishers may make them come across as aggressive, to the detriment of their reputation.

The above conclusion comes with an important qualification. Namely, our studies indicate that the negative effect of showing anger disappears when the norm violation is sufficiently severe (e.g., vandalism or reckless speeding outside a school). We observed a strong relation with the appropriateness of not punishing at all, such that in situations where it was appropriate to refrain from punishment it was particularly inappropriate to show anger, and vice versa.

The appropriateness of refraining from punishment was determined by the severity of the norm violation, but also of your relation to the violator. We found reprimanding a violator to be less appropriate for a stranger than for a friend. This is consistent with previous findings that in a situation where both a friend and a stranger is present, the expectation is for the friend of the violator to speak up rather than the stranger (Strimling & Eriksson, 2014).

We found peer punishment by strangers to be appropriate only for severe norm violations. For mild norm violations, strangers behave more appropriately if the do not reprimand the violator. The latter finding is in line with previous research on judgments of peer punishment in social dilemmas (e.g., Cinyabuguma et al., 2006; Eriksson et al., 2016; Kiyonari & Barclay, 2008), which suggests that non-cooperative behavior in the presumed prototypical social dilemmas used in these studies is generally considered only a mild norm violation. An important conclusion we draw from our study is that from a psychological perspective there can be no such thing as a prototypical social dilemma. Even situations as similar as polluting the common environment with either noise or litter yielded distinct results; in the latter situation it was appropriate for a stranger to reprimand, whereas in the former situation it was more appropriate to refrain from reprimanding.

We also studied the effect of whether violators changed their ways after being reprimanded. Although doing so had a clear positive effect on the violator's approval ratings, it had no effect on the peer punisher's approval ratings. In other words, peer punishers were not judged by the consequences of their acts. Thus, it seems that punishers cannot count on improving their reputations by being effective. This finding of non-consequentialist judgments of peer punishers in these sceanrios adds to previous literature that has focused on judgments of formal punishment (e.g., Baron & Ritov, 1993; Sunstein et al., 1998). As we noted in the introduction, social dilemma researchers tend to theorize about peer punishment based on its beneficial consequences on behavior. Our findings suggest a discrepancy between such theories and the actual psychology of peer punishment. However, we acknowledge that our data are limited to just a few scenarios and that only immediate consequences on behavior were manipulated. The extent of consequentialism in norms about peer punishment deserves further empirical study. A related issue is that if a bystander does not punish a norm violation, there might be someone else who can be counted upon to do it. This will depend on the situation. Such considerations could potentially affect the appropriateness of peer punishment. Whether they do is an open question that we have not examined so far.

In the studies presented here we relied on vignettes, following some previous work on reactions to peer punishers (Strimling & Eriksson, 2014). The factors we examined could be studied also in the field, for instance by extension of the paradigm of Balafoutas et al. (2014). Another limitation is that our samples were all from Western countries (US and Sweden). In ongoing work we have found cross-cultural variation in norms about peer punishment in a social dilemma; it is an open question whether and how the influence of the factors we have studies here may be moderated by culture.

In conclusion, our studies provide insight into how peer punishers can avoid negative reputational effects; for instance, it seems to be more important to refrain from showing anger, and to avoid punishing strangers for mild violations, than to be effective at changing the violator's behavior. Our studies also point to the importance of going beyond economic games and consider genuinely psychological factors when studying peer punishment.

References

Averill, J. R. (1983). Studies on anger and aggression: Implications for theories of emotion. *American Psychologist*, *38*, 1145–1160.

Averill, J. R. (2012). *Anger and aggression: An essay on emotion*. Springer Science & Business Media.

Axelrod, R. M. (1986). An evolutionary approach to norms. *American Political Science Review*, *80*, 1095–1111.

Balafoutas, L., Nikiforakis, N., & Rockenbach, B. (2014). Direct and indirect punishment among strangers in the field. *Proceedings of the National Academy of Sciences*, *111*, 15924–15927.

Balafoutas, L., Nikiforakis, N., & Rockenbach, B. (2016). Altruistic punishment does not increase with the severity of norm violations in the field. *Nature Communications*, *7*, 13327.

Balliet, D., Mulder, L. B., & Van Lange, P. A. (2011). Reward, punishment, and cooperation: a meta-analysis. *Psychological Bulletin*, *137*, 594–615.

Baron, J. (1994). Nonconsequentialist decisions. *Behavioral and Brain Sciences*, *17*, 1–10.

Baron, J., & Ritov, I. (1993). Intuitions about penalties and compensation in the context of tort law. *Journal of Risk and Uncertainty*, *7*, 17–33.

Baron, J., & Ritov, I. (2009). The role of probability of detection in judgments of punishment. *Journal of Legal Analysis*, *1*, 553–590.

Berkowitz, L. (1990). On the formation and regulation of anger and aggression: A cognitive-neoassociationistic analysis. *American Psychologist*, *45*, 494–503.

Brauer, M., & Chaurand, N. (2010). Descriptive norms, prescriptive norms, and social control: An intercultural comparison of people's reactions to uncivil behaviors. *European Journal of Social Psychology*, *40*, 490–499.

Brauer, M., & Chekroun, P. (2005). The relationship between perceived violation of social norms and social control: Situational factors influencing the reaction to deviance. *Journal of Applied Social Psychology*, *35*, 1519–1539.

Chaurand, N., & Brauer, M. (2008). What determines social control? People's reactions to counternormative behaviors in urban environments. *Journal of Applied Social Psychology*, *38*, 1689–1715.

Cinyabuguma, M., Page, T., & Putterman L. (2006). Can second-order punishment deter perverse punishment? *Experimental Economics*, *9*, 265–279.

Dance, A. (2015). News Feature: How online studies are transforming psychology research. *Proceedings of the National Academy of Sciences*, *112*, 14399–14401.

Dodge, K. A., Lochman, J. E., Harnish, J. D., Bates, J. E., & Pettit, G. S. (1997). Reactive and proactive aggression in school children and psychiatrically impaired chronically assaultive youth. *Journal of Abnormal Psychology*, *106*, 37–51.

Elster, J. (1989). *The cement of society: A study of social order*. Cambridge: Cambridge University Press.

Eriksson, K., Andersson, P.A., & Strimling, P. (2016). Moderators of the disapproval of peer punishment. *Group Processes and Intergroup Relations*, *19*, 152–168.

Eriksson, K., Cownden, D., Ehn, M., & Strimling, P. (2014). "Altruistic" and "antisocial" punishers are one and the same. *Review of Behavioral Economics*, *1*, 209–221.

Eriksson, K., Strimling, P., & Ehn, M. (2013). Ubiquity and efficiency of restrictions on informal punishment rights. *Journal of Evolutionary Psychology*, *11*, 17–34.

Eriksson, K., Strimling, P., Andersson, P.A., & Lindholm, T. (2017). Costly punishment in the ultimatum game evokes moral concern, in particular when framed as payoff reduction. *Journal of Experimental Social Psychology*, *69*, 59–64.

Felson, R. B. (1981). An interactionist approach to aggression. In J. T. Tedeschi (Ed.) *Impression management theory and social psychological research*, 181–200. New York: Academic Press.

Fehr, E. & Gächter, S. (2000). Cooperation and punishment in public goods experiments. *American Economic Review*, *90*, 980–994.

Gardner, A., & West, S. A. (2004). Cooperation and punishment, especially in humans. *The American Naturalist*, *164*, 753–764.

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

Henrich, J., & Boyd, R. (2001). Why people punish defectors: Weak conformist transmission can stabilize costly enforcement of norms in cooperative dilemmas. *Journal of Theoretical Biology*, *208*, 79–89.

Kiyonari, T. & Barclay, P. (2008). Cooperation in social dilemmas: Free riding may be thwarted by second-order reward rather than by punishment. *Journal of Personality and Social Psychology*, *95*, 826–842.

Knutson, B. (1996). Facial expressions of emotion influence interpersonal trait inferences. *Journal of Nonverbal Behavior*, *20*, 165–182.

Lochman, J. E., Barry, T., Powell, N., & Young, L. (2010). Anger and aggression. In *Practitioner's guide to empirically based measures of social skills*, pp. 155–166. Springer New York.

Nakao, H., & Machery, E. (2012). The evolution of punishment. *Biology & Philosophy*, *27*, 833–850.

Rouse, S. V. (2015). A reliability analysis of Mechanical Turk data. *Computers in Human Behavior*, *43*, 304–307.

Sabini, J. P., & Silver, M. (1978). Moral reproach and moral action. *Journal for the Theory of Social Behaviour, 8*, 103–123

Strimling, P., & Eriksson, K. (2014). Regulating the regulation: Norms about how people may punish each other. In P. Van Lange, T. Yamagishi & B. Rockenbach (eds.) *Social dilemmas: Punishment and rewards*, pp. 52–69. Oxford University Press, Oxford,

Sunstein, C. R., Kahneman, D., & Schkade, D. (1998). Assessing punitive damages (with notes on cognition and valuation in law). *The Yale Law Journal*, *107*, 2071–2153.

Sunstein, C. R., Schkade, D. A., & Kahneman, D. (2000). Do people want optimal deterrence?. *The Journal of Legal Studies*, *29*, 237–253.

Van Doorn, E. A., Heerdink, M. W., & Van Kleef, G. A. (2012). Emotion and the construal of social situations: Inferences of cooperation versus competition from expressions of anger, happiness, and disappointment. *Cognition & Emotion*, *26*, 442–461.