Moderators of the disapproval of peer punishment

Kimmo Eriksson[1,2], Per A. Andersson[3], Pontus Strimling[3,1]

[1]Stockholm University, Centre for the Study of Cultural Evolution, SE-10691 Stockholm, Sweden

[2]Mälardalen University, School of Education, Culture and Communication, Box 883, SE-72123 Västerås, Sweden

[3]Institute for Analytical Sociology, Linköping University, SE-58183 Linköping, Sweden

Corresponding author: Kimmo Eriksson, Mälardalen University, School of Education, Culture and Communication, Box 883, SE-72123 Västerås, Sweden. Email: kimmo.eriksson@mdh.se Phone: +46 21101533.

Funding

This research was funded by the Swedish Research Council [grant numbers 2009-2390, 2009-2678] and the European Research Council [grant number 324233].

This is a postprint version of: Eriksson, K., Andersson, P. A., & Strimling, P. (2016). Moderators of the disapproval of peer punishment. *Group Processes & Intergroup Relations*, *19*(2), 152-168. https://doi.org/10.1177/1368430215583519

Abstract

Recent studies have found disapproval of peer punishment of norm violations. This seems puzzling, given the potential benefits peer punishers contribute to the group. We suggest part of the answer is that peer punishers tend to come across as aggressive and as such may be viewed as more problematic than beneficial to have around. We used simple computer animations of geometric shapes to enact 15 precise variations of social sanctions against a norm violator. More than 1,800 subjects were recruited to watch an animation and judge the behavior and character of the animated agents. They also completed a trait aggression measure. Across the variations peer punishment was typically disapproved of, especially when severe or openly aggressive, and especially by subjects low on trait aggression. We conclude that there seems to be a social norm against peer punishment and that dislike of aggressiveness seems to be part of the reason why.

Keywords. trait aggression; social norms; social judgment; social control; peer punishment

Introduction

Consider a group of peers. One group member, Viola, violates a social norm. Another group member, Punisha, reacts to Viola's norm violation by some kind of informal punishment. Other group members also tend to disapprove of Viola's behavior. The question is, will they therefore approve of Punisha punishing Viola?

This question is motivated by the literature on cooperation in situations where there is a conflict between self-interest and group-interest. In such "social dilemmas", economic experiments demonstrate that groups may achieve substantially higher levels of cooperation if it is in the power of group members to punish each other (e.g., Yamagishi, 1986; Fehr & Gächter, 2000; Balliet, Mulder, & Van Lange, 2011). In this research tradition it is common to regard informal punishment of selfish behavior as a public good in itself. Consequently, those who do not punish free-riders among their peers are regarded as free-riding on others' punishment efforts. Because of the great potential benefits from cooperation it has been argued that biological and cultural evolutionary processes will shape social behavior such that all free-riding, including *not* punishing free-riders, is subject to peer punishment (e.g., Henrich & Boyd 2001). From this popular hypothesis it follows that other group members should definitely tend to approve of Punisha in the above scenario. Indeed, had Punisha not subjected Viola to punishment, we should expect other peers to want to punish both Viola for violating the norm *and* Punisha for not punishing Viola.

Some experiments on economic games have addressed this prediction by letting participants react to peer punishment, either by participating in a second stage of sanctions (e.g., Cinyabuguma et al., 2006; Kiyonari & Barclay, 2008) or by expressing who they would prefer to be partnered with in subsequent games (Horita, 2010). Findings from such studies were reviewed by Strimling and Eriksson (2014), who concluded that the prediction is generally not supported by the data. Reactions to use of peer punishment in economic games actually tend to be negative. However, there is general concern about the external validity of economic laboratory games (Levitt & List, 2007). In particular, economic experiments on peer punishment implement it in terms of reducing the monetary payoff of another participant. Such material sanctions between peers are arguably uncommon in most real-life situations. The social psychological literature on informal sanctions between peers instead focuses on social forms of sanctions. For instance, in a field experiment where confederates violated various prosocial norms, researchers categorized informal sanctions as angry looks, loud audible sighs, loud comments to oneself, and comments to the norm violator ranging from polite to aggressive personal insults (Brauer & Chekroun, 2005). Unfortunately, such field

experiments tend not to study how people (other than the norm violator) react to such social sanctions.

Thus, there is very limited research on how peer punishment of norm violators is viewed in real life. Recently this question has been approached using cross-cultural online surveys (Eriksson, Strimling & Ehn, 2013; Strimling & Eriksson, 2014). Consistent with the findings in economic games, survey respondents tended to disapprove of peer punishers. This finding held across cultures and across a range of situations. Strimling and Eriksson (2014) drew the conclusion that use of informal punishment between peers tend to be socially proscribed rather than prescribed. This poses a theoretical puzzle. Why is peer punishment disapproved of, despite its potential to solve the problem of achieving cooperation with all the benefits that entails? Below we develop a hypothesis to explain this puzzle.

Is peer punishment condemned as aggression?

In their most important study about peer punishment, Strimling and Eriksson (2014, Study 3) used four scenarios describing how one person acted against the interests of others in the group. Every scenario described a non-punisher and a peer punisher. Whereas the former was described as simply letting the selfish behavior go, the latter yelled about the selfish behavior (in three scenarios) or used a monetary punishment (in a scenario describing an economic experiment). Respondents from United States and India compared the two peers on seven traits. Results were consistent across all scenarios (including both yelling and monetary punishment) and both countries: Compared to the non-punisher, the peer punisher was judged as *less* preferable to spend time with; *less* likely to adhere to standard norms of behavior, *less* likely to take others' interests into account, and *less* likely to be trustworthy; but *more* likely to punish other people unfairly, *more* likely to be an angry person, and *more* likely to create bad morale in the group.

Our interpretation is that use of peer punishment may be taken as a sign that the punisher has an angry disposition and does not have the group's interest in mind. Various results from previous studies are consistent with this interpretation. Use of peer punishment in economic lab experiments seems generally not motivated by the group outcome (Eriksson, Cownden, Ehn & Strimling, 2014). Experiments on the ultimatum game indicate that rejections of low offers, which is often conceived of as a form of peer punishment, is related to high testosterone levels (Burnham, 2007) and motives of assertion (Yamagishi et al., 2012). Field studies indicate that punishment between peers is often driven by hostile emotions

(Chaurand & Brauer, 2008). Thus, a tendency to judge peer punishers as angry and not having the group's interest in mind may be well-motivated.

Note that disapproval of peer punishment is still puzzling from a rational point of view. After all, even if the punisher is motivated by anger rather than the good of the group, the group could still benefit from having a member who is prepared to punish any peers that behave selfishly.

We think the puzzle dissolves of one takes into account the bigger picture of *all* situations that occur in a group. An angry group member may well be valuable in a particular situation and still be a problem in other situations where the anger is not aligned with the interest of the group. Indeed, it has recently been argued that bullies have been a major problem for groups to deal with in human evolutionary time (Sterelny, 2014). In line with this notion, a long-term trend towards ever stronger norms against physical aggression has been documented (Pinker, 2011). Even aggression towards an outgroup member is not particularly popular (Vandello, Ransom, Hettinger, & Askew, 2009). Several studies have found that a desire to punish someone tends not to be considered a good justification for aggression (Lagerspetz & Westman,1980; Ramirez, 1991).

Our main hypothesis is that if a peer punisher comes across as aggressive then it does not help that the aggression is directed towards a behavior that everyone disapproves of. The positive aspect of the fact that a bad behavior was punished will tend to be trumped by the negative impression of perceived aggressiveness.

Research questions about judgments of peer punishers and the act of peer punishment

The main hypothesis motivates a number of questions. The first question asks what is included in the negative impression created by use of aggressive peer punishment. As described above, previous research suggests that when someone yells at a norm violator, or uses monetary punishment, it tends to be taken as an indication that the peer punisher has a number of undesirable traits (Strimling & Eriksson, 2014). Because the act of punishing may be beneficial to the group, it could be argued that even if people are wary of the person who performs an act of punishment it may still be viewed as appropriate and the punisher may still be viewed as an asset to the group. Thus, we are interested in establishing whether the negative impression of peer punishers extends to the act of punishment being viewed as *inappropriate* and the peer punisher being viewed as a *problem* to the group rather than an asset. The effect we expect is that people's general disapproval of peer punishment will manifest in all these aspects.

The main effect that peer punishment leads to general disapproval may be moderated by several factors. One factor is the characteristics of the individual that makes the judgment. It is known that approval of aggressive behavior is linked to own level of aggression (Huesmann & Guerra, 1997). From the main hypothesis that peer punishers tend to be viewed negatively because they come across as aggressive, we therefore expect less disapproval of peer punishers among people whose own level of aggression is high.

Other factors relate to variations in the way the punishment is carried out. For instance, a punisher may act alone or have backing from less active supporters. When a punisher has supporters they should come across as less aggressive than the actual punisher. They are therefore expected to be less negatively evaluated. However it is unclear whether the evaluation of the actual punisher should depend on whether he or she is acting alone or with backing from supporters. On the one hand, the presence of supporters may make the punishment seem more in line with social norms. On the other hand, the presence of supporters may give the impression of a group ganging up on a lonely victim.

Punishments may also vary in their severity. Game theoretic accounts of punishment assume that the more severe they are, the better they will serve the group as deterrents of future bad behavior (e.g., Becker, 1968). From that point of view, more severe punishments should gain higher approval. On the other hand, more severe punishment should come across as more aggressive. Based on the hypothesis that perceived aggressiveness trumps the potential benefits of punishment, we expect more severe punishment of a peer to lead to more negative judgments of the punisher.

There are also different kinds of punishment. The above-mentioned economic laboratory games use an economic kind of peer punishment, such that the punished party's monetary payoff is reduced. Experiments on aggressive behavior in social psychology have often used physical kinds of peer punishment, such as administration of hot sauce, sound blasts, or electric shocks. Whether an economic or physical kind of punishment will be felt as the most severe by the punished party is likely to be subjective. However, we expect economic punishments to be generally perceived as less aggressive than physical punishments. The main hypothesis then predicts that an economic peer punisher will tend to be judged less negatively than a physical peer punisher.

In real life it is common that reactions to a norm violation are neither economic nor physical, but rather take the form of a simple verbal confrontation. As a verbal confrontation may serve to establish a norm without causing any harm it should be regarded as more appropriate than economic or physical punishments and the verbal punisher should be less

likely to be viewed as a problem for the group. However, verbal confrontations still tend to be driven by hostile emotions (Chaurand & Brauer, 2008) and may therefore come across as aggressive. Indeed, the results from the yelling scenarios of Strimling and Eriksson (2014) indicate that also verbal punishers may be evaluated negatively.

Outline of studies

To obtain stimulus material representing a norm violation and different reactions to the norm violation, we constructed abstract animations of a group of geometric shapes. A classic finding in psychological research is that certain patterns of movements of abstract shapes are perceived as having social and emotional content (Heider & Simmel, 1944; Rimé et al., 1985). Such animations have later been used with success in studies of social attribution (Bell, Fiszdon, Greig, Wexler, 2010; Castelli, Happé, Frith, Frith, 2000; Congiu, Schlottmann, Ray, 2009; Klin, 2000). In a novel application of the technique of geometric animations, we here use it to study social judgment. Note that the most closely related previous research on social judgment of peer punishers used vignettes specifying people in a certain real-world context (Strimling & Eriksson, 2014). Geometric animations provide a simple way to enact a large set of precise variations of a peer punishment scenario. Because animations do not rely on verbal descriptions but on direct viewing, we expect them to be more salient than vignettes and less subject to "reading between the lines". Further, they can be used across cultures without translation.

Four studies were conducted. In every study, participants viewed animations and made social judgments of the animated agents. Participants' level of trait aggression was also measured. The first two studies (with American and Swedish participants, respectively) used scenarios with physical punishment of varying severity and it also varied whether the punisher acted alone or had supporters. In a control scenario there was no peer punishment at all. The third study examined economic punishment in the same manner. The last study examined verbal confrontation in the same manner and also tied up some loose ends from the previous studies.

We report for each study how we determined our sample size, all data exclusions (if any), all manipulations, and all measures.

Study 1: Physical sanctions

In the first study we investigated American participants' reactions to physical sanctions of the norm violator.

Method

All studies employed animations of geometric shapes to represent behavioral variations in a scenario about sharing of a common resource. The animations can be accessed at www.pontusstrimling.com/animations/

In dramatic terms the basic scenario can roughly be divided into three scenes: The first scene establishes a norm: agents take turns to harvest a common resource. In the second scene, one of the agents violates this norm of turn-taking by harvesting the entire remaining resource. The third scene shows how the other agents acted to sanction the norm violation which varied depending on the condition.

Specifically, the animation had the following building blocks (with the intended interpretation presented within parentheses): a white stage on which the action takes place; four triangles (the agents) of different colors, initially positioned in quadratic structures at each of the four corners of the stage; and a number of small circles (resource units), initially aggregated in the center. As the animation started each triangle moved, one by one, to the center to move one circle from there to their respective corner (harvesting the common resource). Figure 1 illustrates the overall features of the animations, the positioning of the triangles, and shows the blue triangle returning to its corner with a newly harvested circle. During the first scene each triangle went twice to the center to harvest a circle, increasing the number of circles at their respective corners while decreasing the number of circles at the center. At this point the purple triangle approached the center, first turning to the left and to the right (sneaking and looking around), and then proceeding to push all the remaining circles back to its corner. This concluded the norm violation scene. From this point onwards there were a number of variations of the animation showing how the norm violation was sanctioned.

Conditions

Sanctions could be either collective, individual, or absent. In the *collective sanction* type of variation the green triangle moved to the center and turned around (looking for the circles). It then moved first to the pink triangle and then to the blue triangle (urging them to come), who both followed it back to the center. At the center they turned back and forth to each other (looking around and discussing the situation) before they together moved to the norm violator's corner. They gathered outside this corner (seeing the stolen resources there) before turning away and forming a semi-circle in which they turned back and forth to each other

(having a meeting). At this point one of several punishments followed, to be described in the studies.

In the *individual sanction* type of variation, the green triangle moved to the center, turned around, and then returned back to its corner (letting the norm violation go). Following this, the blue triangle moved to the center, turned left and right (looking around), and then moved on its own to the corner of the purple triangle (confronting the norm violator). At this point followed one of several punishment conditions described below.

The *no sanction* condition started identically as the individual sanctions, but instead of the blue triangle moving from the center to confront the norm violator it turned around and moved back to its own corner (letting the norm violation go), and the animation ended.

Study 1 used a between-subjects design with 100 participants in each of five conditions: no sanction (NO), collective weak physical sanction (CWP), individual weak physical sanction (IWP), collective strong physical sanction (CSP), and individual strong physical sanction (ISP).

In the collective sanctions (CWP and CSP) conditions, the green and pink triangles aligned and pointed at the third one, the blue triangle, with their tips (encouraging it to act as the punisher). They then moved to face the norm violator in its corner, where the blue triangle, now in the middle of the three, delivered the sanction. Following the sanction, the triangles all returned to their respective corners, and the animation ended. In the individual sanctions (IWP and ISP) conditions, the blue triangle was on its own when it moved to the corner and delivered the sanction.

A weak physical sanction (CWP and IWP) was implemented by the blue triangle making two moves against the purple triangle (shoving it backwards) with no damage being done. A strong physical sanction (CSP and ISP) was implemented by the blue triangle attacking the purple triangle four times until it shattered (hurtful engagement).

Participants

Five hundred participants (43% female, age ranging from 17 to 76 years with median 31 years) were recruited among American users of Amazon Mechanical Turk (mturk.com) to take part in an online experiment. Participation was rewarded by a fee of 1 US dollar. The sample size (100 subjects per condition) was determined by convenience and set in advance.

Procedure

Subjects were instructed that they were going to watch a short (1-2 minutes) animation of triangles and report how they felt about the behavior of the different triangles. Animations were shown as embedded YouTube movies, and subjects were instructed that they could replay them if they wished. All questions about the animation were given on the same screen, so subjects could watch the animation again if they felt they needed it to answer the questions. Subjects were told that triangles would be referred to by their color names. To eliminate any ambiguity, they were explicitly told which color name (Blue, Green, Pink, or Purple) referred to the triangle in which corner.

Throughout most of the questionnaire, responses were given on a seven point Likert scale from 1=Strongly Disagree to 7=Strongly Agree. Subjects rated each of the four triangles on the same three items: whether the triangle's behavior was *appropriate;* whether they would like to *spend time* with a person who behaves like that triangle; and whether they would consider a person who belonged to their group and behaved like that triangle to be *a problem (rather than an asset) for the group.*

Three items dealt with subjects' experience of rating the animations: whether they were *confident they had judged the correct triangles;* whether triangles had looked *as if they were "alive"*; and whether the triangles' motion looked *as if it was goal-directed and intentional*.

Five additional items focused on the blue triangle, asking for subjects' spontaneous interpretation of the blue triangle's character. All items started *I think BLUE is someone who...*, and then continued: *is generally trustworthy; is generally angry; takes others' interests into account*; *would punish others unfairly*; and *generally follows standard norms of behavior.* These items were adapted from a previous study of punishers (Strimling & Eriksson, 2014). The response scale was "no", "don't know", and "yes" (coded as 0, 1, and 2).

Subjects were then asked how many times they had watched the animation and given the opportunity to motivate their judgments of the animation in free text. They then completed a short Trait Aggression scale (Bryant & Smith, 2001). Although not reported below, the study also included a short Big Five scale (Gosling, Rentfrow, & Swann, 2003) and either a Social Dominance Orientation scale (Pratto et al., 1994) or a Justice-Vengeance Scale (Ho et al., 2002).

Results

We used an alpha level of .05 for all statistical tests.

Validity of animations

Before addressing our research questions we analyzed the validity of the animations as a method to elicit social judgments of norm violations and sanctions. Subjects were confident they had judged the correct triangles (median response 7 out of 7). They also thought the triangles looked alive (median response 6 out of 7) and that their movement was goal-directed and intentional (median response 7 out of 7). Purple, the "norm violator", was typically rated as being a problem for the group (median response 7 out of 7) and not behaving appropriately (median response 1 out of 7). Thus, as desired, Purple's action was perceived as a clear norm violation.

Most subjects (81%) wrote a motivation of their judgments, and these motivations typically indicated vivid impressions of triangles' intentions, emotions and morality. Here are two random examples of motivations given in different conditions:

> "I think Blue has a group's best interests at heart, and I think he stands for fairness and is willing to fight for equality. Green and Pink, it seems like they want equality, but rather have someone fight on their behalf rather than stand up for their own selves and fight for it. It seems to me that they talked Blue into fighting the Purple triangle who was completely out of line for stealing all the red balls." (CWP)
>
> "Blue had some serious anger issues!! Purple took way too many dots. Pink and green played fairly." (ISP)

In sum, we conclude that animations of geometric shapes seem to be a valid way of tapping into people's social judgments.

Judgments of the peer punisher

Subjects made a total of eight judgments of Blue, the triangle that was the active punisher in all conditions involving peer punishment. Table 1 presents mean values for each item in each condition.

A factor analysis strongly supported a single factor explaining 64% of the variance, on which the five positive items loaded positively and the three negative items (i.e., is a problem for the group, is generally angry, and would punish others unfairly) loaded negatively. After reverse-coding the latter three items a z-score transformation was performed for each item. The average of the eight z-scores will be referred to as the *Blue approval index* (Cronbach's alpha = .92). Mean values of the Blue approval index, per condition, can be found in Table 1.

Approval of non-punishment vs. peer punishment

The results in Table 1 indicate a robust difference in Blue approval between non-punishment and peer punishment: On every single item, Blue gained higher mean approval in the no sanction condition than in any of the peer punishment conditions. The effect size was generally very large; the difference in Blue approval between the no sanction condition (M = 0.79) and the pooled peer punishment conditions (M = -0.20) represents a Cohen's d of 1.24.

Moderating effects of support and severity on approval of peer punishment

Now consider how the Blue approval index varied within the set of peer punishment conditions. These conditions followed a two-by-two design: two levels of punishment severity (weak, strong) and two levels of support of the punisher (individual, collective). A two-way analysis of variance yielded a large main effect of severity, $F(1, 396) = 99.93$, $p < .001$, such that the average approval was lower for strong punishment (M = -0.53, SD = 0.63) than for weak punishment (M = 0.13, SD = 0.71), d = 0.89. There was also a small main effect of support for the punisher, $F(1, 396) = 8.24$, $p = .004$, such that approval was lower for individual punishment (M = -0.29, SD = 0.75) than for collective punishment (M = -0.10, SD = 0.75), d = 0.25. The interaction effect was non-significant, $F(1, 396) = 0.09$, $p = .76$.

Ratings of other group members

We now turn to ratings of the remaining group members, Green and Pink. Mean ratings on the three items that were asked about each of these triangles are reported in Table 1. After reverse-coding the problem-for-the-group item, a z-score transformation was performed for each item. The average of the six z-scores will be referred to as the *Green-Pink approval index* (Cronbach's alpha = .82). Mean values of the Green-Pink approval index per condition can be found in Table 1.

In a post hoc analysis, two things stand out as noteworthy. First, Green-Pink approval was higher in the no sanction condition (M = 0.33) than in any of the peer punishment conditions (pooled M = -0.08), d = 0.57. Thus, it seems that Blue's deployment of peer punishment reflected badly also on the group as a whole. This is consistent with prior findings that people tend to make poorer judgments of a group if one of its members punishes a peer (Strimling & Eriksson, 2014, Study 4).

Second, Green-Pink approval was particularly low in the collective strong punishment condition (M = -0.42). Thus, participants tended to disapprove of active supporters of strong peer punishment.

Trait aggression and approval of peer punishment relative to non-punishment

Responses to the 12-item trait aggression scale (Cronbach's alpha = .85) were averaged to a TA score between 1 and 7 (M = 3.36, SD = 1.05). To measure individual subjects' approval of peer punishment relative to non-punishment we focused on the individual sanction conditions, where we computed the difference between the Blue approval index and the Green-Pink approval index. To examine whether raters' level of trait aggression predicted their approval of peer punishment relative to non-punishment, we calculated the correlation between the approval index difference and the TA score in these conditions. The correlation was positive both in the individual weak sanction condition (r = .28, p = .005, N = 100) and the individual strong sanction condition (r = .44, p < .001, N = 100). Thus, raters that scored higher on trait aggression tended to show less disapproval of punishment relative to non-punishment.

Discussion

A previous study (Strimling and Eriksson, 2014) presented respondents with verbal descriptions of scenarios where one peer reacted to a norm violation by letting it go and another peer reacted by punishing the norm violator. In that study, peer punishers were consistently judged as having a worse character than the non-punisher: less trustworthy, less likely to take others' interests into account, etc. Here we replicated these findings using animations of geometric shapes instead of verbal scenarios and using a between-subjects instead of within-subjects design.

This study was designed to answer several additional research questions as well. First, we found that the disapproval of peer punishment seems to extend beyond judgments of the individual's character to include viewing the individual as a problem for the group and the behavior as inappropriate. Second, we found disapproval of peer punishment to be moderated by several factors. The rater's level of trait aggression predicted less disapproval of peer punishment relative to non-punishment. Further, more severe peer punishment tended to be disapproved of more. All of these findings are consistent with our hypothesis that disapproval of peer punishment is related to it being viewed as aggressive.

We also compared a peer punisher acting alone with a peer punisher backed up by supporters. The presence of supporters made the peer punisher look somewhat better (but at the same time it made the supporters look worse).

Study 2: Replication of Study 1 under lab conditions

Study 1 used an on-line sample recruited among American users of the Amazon Mechanical Turk. It is an important test of robustness to establish that the same findings are obtained also under other circumstances. We therefore replicated Study 1 in a Swedish computer-based laboratory for social psychology experiments.

Method

Participants

An invitation to participate was sent by email to students at a Swedish university who had previously expressed interest in participating in experiments. 162 participants were recruited (45% female, 51% male, 3% unknown, age ranging from 18 to 74 years with median 25 years). Participation in a session was rewarded by a show-up fee of 60 Swedish kronor (about 10 US dollars). Sessions included both this study and other unrelated studies for a total duration of about one hour. The sample size was set to be smaller than in Study 1 due to the higher cost of participation fees.

Procedure

Subjects showed up at the laboratory at the start of a session and were led to cubicles separated by screens. Every cubicle had a desk with a computer on which the experiment was run. The software randomly assigned subjects to one of the five conditions, the same as in Study 1, and presented subjects with the same battery of questions (in Swedish translation), including the Big 5 and Social Dominance Orientation scales that we do not report.

Results

The results of this study generally replicated the findings of Study 1, although statistically weaker because of the considerably smaller sample size. First we calculated the Blue approval index (Cronbach's alpha = .89). Mean values per condition are reported in Table 2. The difference in Blue approval between the no sanction condition (M = 0.71) and the pooled peer punishment conditions (M = -0.17) represents a Cohen's d of 1.18.

A two-way analysis of variance revealed a large main effect of severity, $F(1, 126) = 24.89$, $p < .001$, with average approval lower for strong punishment (M = -0.46, SD = 0.61) than for weak punishment (M = 0.10, SD = 0.66), d = 0.80. The main effect of support for the punisher was not statistically significant, $F(1, 126) = 1.66$, $p = .20$, but resembled Study 1 in

terms of effect size such that approval tended to be lower for individual punishment (M = -0.24, SD = 0.69) than for collective punishment (M = -0.11, SD = 0.70), d = 0.20. The interaction effect was non-significant, $F(1, 126) = 0.05$, $p = .82$.

Descriptive statistics of the Green-Pink approval index (Cronbach's alpha = .90) are presented in Table 2. Just as in Study 1, approval was highest in the no sanction condition and lowest in the collective strong punishment condition.

We then examined the trait aggression scale (Cronbach's alpha = .85, M = 3.31, SD = 0.75). In line with Study 1 the correlation between the TA score and the difference between the Blue and Green-Pink approval indexes was positive both in the individual weak sanction condition (r = .31, p = .07, N = 33) and the individual strong sanction condition (r = .40, p = .024, N = 31).

Discussion

The findings from the American on-line sample of Study 1 was replicated using the same animations but in a study conducted in a lab environment with Swedish students. It seems to be a general observation that findings tend to generalize across lab and on-line studies (e.g., Buhrmester, Kwang & Gosling, 2011; Paolacci, Chandler & Ipeirotis, 2010). It is more noteworthy that the findings generalized across the cultural gap between Americans and Swedes. This generalizability is consistent with previous findings of cross-cultural similarities in the view of informal punishment (Strimling & Eriksson, 2014).

Study 3: Economic sanctions

So far we have found that "physical" sanctions of a norm violator tend to be disapproved of. Here we investigate "economic" sanctions instead, in terms of resources being taken from the norm violator.

Method

Participants

Five hundred participants (47% female, age ranging from 18 to 75 years with median 30 years) were recruited among American users of Amazon Mechanical Turk (mturk.com) to take part in an online experiment. Participation was rewarded by a fee of 1 US dollar. The sample size was set to match that of Study 1.

Animation conditions

The experiment used a between-subjects design with 100 participants in each of five conditions: no sanction (NO), collective weak economic sanction (CWE), individual weak economic sanction (IWE), collective strong economic sanction (CSE), and individual strong economic sanction (ISE).

In the collective sanctions (CWE and CSE) conditions, the three triangles moved together to the corner where the norm violator was located; each of them took the same number of the norm violator's circles and then returned to their respective corners (at which point the animation ended). In the individual sanctions (IWE and ISE) conditions, the blue triangle was on its own and took a number of the norm violator's circles, left two thirds in the center and brought the remaining third to its own corner (at which point the animation ended).

A weak economic sanction (CWE and IWE) was implemented by the purple triangle, the "norm violator" being deprived only of the circles it had taken out of turn (leaving it with those resources that had been harvested according to the norm). A strong economic sanction was implemented by the purple triangle being deprived of all its circles (leaving the norm violator with no resources at all).

Procedure

The procedure followed Study 1 with one minor change: For the five items focusing on subjects' spontaneous interpretation of the blue triangle's character, the response scale was changed to the same 7-point Likert scale used throughout the rest of the questionnaire. Although not reported below, the study also included the same short Big Five scale and Justice-Vengeance Scale as Study 1.

Results

The analysis follows that of Study 1. As in Study 1 there was high agreement that the triangles looked alive (median response 6) and that their movement was goal-directed and intentional (median response 7).

Descriptive statistics, per condition, for the Blue and Green-Pink approval indices (Cronbach's alpha = .88 and .79, respectively) are reported in Table 3. The difference in Blue approval between the no sanction condition (M = 0.45) and the pooled peer punishment conditions (M = -0.11) represents a Cohen's d of 0.75. This is a large effect but not as large as for the physical sanctions in Studies 1 and 2.

A two-way analysis of variance of Blue approval revealed a medium-sized main effect of severity, $F(1, 396) = 23.55$, $p < .001$, with average Blue approval lower for strong punishment ($M = -0.29$, $SD = 0.81$) than for weak punishment ($M = 0.07$, $SD = 0.66$), $d = 0.47$. There was a small main effect of support for the punisher, $F(1, 396) = 6.17$, $p = .013$, such that approval tended to be lower for individual punishment ($M = -0.20$, $SD = 0.83$) than for collective punishment ($M = -0.02$, $SD = 0.66$), $d = 0.24$. The interaction effect was non-significant, $F(1, 396) = 0.01$, $p = .91$.

Also Green-Pink approval followed the same pattern as in Studies 1 and 2: Approval was highest in the no sanction condition and lowest in the collective strong punishment condition.

We then examined the trait aggression scale ($M = 3.39$, $SD = 0.98$). The correlation between the TA score and the difference between the Blue and Green-Pink approval indexes showed a very weak and non-significant positive tendency in the individual weak sanction condition ($r = .08$, $p = .41$, $N = 100$) but was statistically significant in the individual strong sanction condition ($r = .24$, $p = .017$, $N = 100$).

Discussion

The results of this study indicate that the social norm against peer punishment is not restricted to physical sanctions. Also economic sanctions, in the form of taking back resources that a norm violator has taken out of turn from a common pool, received less approval than letting the norm violation go. We also replicated the other main findings from Studies 1 and 2: First, more severe punishment was less approved of. Second, collective punishment was more approved of than individual punishment. Third, when one group member punished a peer it reflected badly even on group members that were not involved in the punishment. Fourth, trait aggression was related to lower disapproval of peer punishment relative to non-punishment.

Table 3 includes ratings of Blue on those items where the same response scale was used as in Study 1. A comparison between the absolute levels of ratings of Blue in Tables 1 and 3 suggests that economic punishments were met with less disapproval than physical punishments. In particular it is noteworthy that the weak physical punishment (shoving Purple once with no damage being done) tended to be more disapproved of than strong economic punishment (depriving Purple of all its resources). Arguably, the latter punishment does more harm. This pattern is consistent with perceived aggression, rather than actual harm done, being the main driver of disapproval.

Study 4: More variations on sanctions

One concern with the previous studies is that the individual sanction conditions all started with Green letting the norm violation go before Blue made the decision to punish. It is possible that Green's behavior established a norm of non-punishment and that Blue's decision to punish would otherwise have been met with approval. In order to test this alternative explanation of Blue's low approval ratings we conducted a new study of individual sanctions in which Blue was always the first one to notice the norm violation. We made this adaption to three conditions from the earlier studies: no sanction, individual weak physical sanction, and individual strong economic sanction.

Another concern with the economic sanctions in Study 3 is that Blue may have come across as selfish by keeping for itself a third of the resources taken from Purple. Thus disapproval might have been driven by the perception of selfishness rather than disapproval of the peer punishment itself. In order to test this alternative explanation we included a new version of individual strong economic sanction in which Blue left all the resources in the middle.

Whereas physical and economic punishments are often used in lab experiments, we discussed in the introduction that real-life social sanctions may typically consist of verbal confrontations instead. To test the approval of verbal confrontations within our animation paradigm we adapted the weak physical punishment condition into a verbal condition by replacing the shove with a speech balloon containing an exclamation mark.

Finally, it seems plausible that approval of peer punishment may depend on whether it works as an effective deterrent of the norm violation. We tested this by adding a condition in which the verbal confrontation was followed by Purple returning the ill-gotten resources to the common pool.

Method

Participants

Six hundred and forty participants (48% female, age ranging from 18 to 73 years with median 30 years) were recruited among American users of Amazon Mechanical Turk (mturk.com) to take part in an online experiment. Participation was rewarded by a fee of 0.50 US dollars. The sample size was set to match that of Studies 1 and 3, i.e., 100 subjects per condition in six conditions, but due to a typo one condition was assigned 140 subjects instead.

Animation conditions

The experiment used a between-subjects design with six conditions: no sanction (NO-B), weak physical sanction (IWP-B), strong economic sanction (ISE-B), strong economic unselfish sanction (ISE-un), verbal sanction (IV), and verbal effective sanction (IVE). There were 100 subjects per condition with the exception of the no sanction condition, which was taken by 140 subjects.

For this experiment the blue triangle, who also carried out the punishments, was the first and only triangle to move following the norm violation by purple triangle. The weak physical sanction (IWP-B) otherwise played out as previously (in Study 1 and 2), with the blue triangle making two moves against the purple triangle, with no damage being done, and then returning to its corner.

Beside the blue triangle being the only triangle to act following the norm violation by purple, the strong economic sanction (ISE-B) also played out as previously (in Study 3) with blue triangle taking all of the norm violator's circles, leaving two thirds in the center and bringing the remaining third into its own corner. In the strong economic unselfish sanction (ISE-un) all actions were identical except that the blue triangle now returned all the circles to the center and did not bring any to its own corner.

The verbal sanctions (IV and IVE) mimicked the physical punishment condition (IWP-B), but with no pushing. Instead, the blue triangle faced the purple triangle and wiggled rapidly side to side while a speech balloon appeared above it with a pulsating exclamation mark (talking). At this the purple triangle backed away slightly. The verbal sanction (IV) ended with the blue triangle returning to its corner. In the verbal effective sanction (IVE) the purple triangle then returned the circles to the middle, thus restoring the situation as it was before the norm transgression.

Procedure

The procedure followed Study 3 with the exception that the Big Five and Justice-Vengeance scales were not included.

Results

We focus the analysis on comparisons of the Blue approval index (Cronbach's alpha = .88) between conditions. Mean values per condition are reported in Table 4.

The effect of having Blue go first

The first concern was whether Blue's decision to punish would have been met with approval if only Green had not established a norm of non-punishment. The present study indicates this is not the case: Both the weak physical and strong economic sanctions remained at considerably lower levels of approval than non-punishment, Cohen's $d = 1.31$ and 0.92, respectively.

Profitable vs. unselfish economic sanctions

The second concern was whether disapproval of economic sanctions was driven by their being profitable and therefore coming across as selfish. A comparison between conditions ISE-B and ISE-un shows that approval ratings rose considerably when Blue did not profit from the economic sanction, $p < .001$, Cohen's $d = 0.70$. Indeed, although mean approval of the unselfish version of economic punishment was lower than for non-punishment, the difference was small and did not quite reach statistical significance, $p = .066$, Cohen's $d = 0.25$.

Verbal sanctions

Verbal sanction (IV) was less approved of than non-sanction, $p < .001$, Cohen's $d = 0.46$, and this held even when the verbal sanction was shown to be effective at changing the norm violator's ways (IV-E), $p = .020$, Cohen's $d = 0.30$. Indeed, there was no statistically significant improvement of approval of verbal sanction when it was shown to be effective, $p = .26$, Cohen's $d = 0.16$.

Discussion

As in the previous three studies, the no sanction condition yielded the highest level of Blue approval. There was a clear difference in approval between non-punishment and all peer punishment conditions with the exception of the unselfish economic punishment. Note that the latter kind of sanction was explicitly prosocial. Such explicitly prosocial sanctions are unlikely to be available in many real-life situations; it seems to be particular to scenarios where a sanction can take the form of restoring a material resource.

General discussion

The topic of this paper is how people view peer enforcement of social norms. In the introduction we reviewed recent empirical work suggesting that peer punishment tends not to be approved of. From a functional perspective this seemed puzzling as peer punishers can promote cooperation that benefits the group. We hypothesized that the answer to this puzzle is that peer punishers will often come across as aggressive, and that the impression of within-group aggression is no negative that it outweighs the potential benefits to the group in the situation at hand.

To examine this hypothesis we conducted a series of experiments where subjects rated animations showing a norm violation and various reactions to it. The target of peer punishment was a behavior that clearly harmed the group. Nonetheless, peer punishers tended to be viewed as behaving less appropriately than a non-punisher and as being more of a problem for the group. Several findings about moderators spoke to the role of perceived aggressiveness in driving disapproval of peer punishment. First, more aggressive raters tended to show less disapproval of peer punishment. Second, more severe punishments tended to be more disapproved of. Third, clearly aggressive but non-harmful punishment tended to be more disapproved of than clearly harmful but material punishment (Study 4).

A limitation of our studies is that only one kind of norm violation was used. When one individual takes more than its share of a common resource it has negative consequences for everyone else in the group. Other norm violations may be neutral in their consequences for others, such as having sex with someone that society does not approve of your having sex with. Norm violations may even be objectively beneficial for everyone else in the group, such as giving more generously to a common cause than other group members are comfortable with (Parks & Stone, 2010). It is an intriguing question whether the view of peer punishment depends on what kind of norm violation it targets. This should be addressed in future research. We note that the animation technique employed in this paper could be used to enact other norm violations as well.

Another limitation of our studies is that all sanctions were direct confrontations. Informal sanctions can also be non-confrontational, such as gossip and avoidance. Non-confrontational sanctions are likely to come across as non-aggressive. Our hypothesis would therefore expect use of non-confrontational sanctions between peers to gain higher social approval. This could also straightforwardly be investigated using the same basic methodology we have used here.

Conclusion

To conclude, when someone violates a prosocial norm it seems to be counter-normative for a peer to react with confrontational punishment even though it may be good for the group. Dislike of aggressiveness seems to be part of the explanation why confrontational peer punishment is frowned upon. However, there are likely to be other contributing mechanisms as well. For instance, it seems very common that peer punishers experience counter-punishment from their target (Balafoutas & Nikiforakis, 2012; Nikiforakis, 2008; Nugier et al., 2007). This experience could, we speculate, contribute to internalization of a norm that you should not punish your peers.

A peer punisher should, it seems, expect to be met with counter-punishment as well as general disapproval. For some people and in some situations, these social costs may be outweighed by the emotional reward of acting out against someone whose behavior is unacceptable. The same kind of emotional reward might be attainable through consumption of popular stories of heroic vigilantes, like Robin Hood or Batman, punishing bad guys. Note, however, that such stories differ from our notion of peer punishment in two critical respects. First, the villains are usually not peers but extraordinarily evil and powerful. Second, the heroes are often mainly focused on preventing evil deeds, or setting them right, rather than punishing them. We would welcome research into how these aspects (the power of the norm violator and the difference between punishment and prevention) affect views of peer reactions to norm violations.

Finally, note that people do not disapprove of punishment in general. Political parties that promise to enforce laws that punish wrongdoers tend to be popular. There is no popular demand for abolishing all punishment from the criminal code. In short, formal punishment is generally approved of. In those situations where punishment could play a positive role and formal punishment is not available, it would be desirable if peer punishment could gain general approval. If, as we hypothesize, a key issue with peer punishers is that they tend to come across as aggressive, then genuinely prosocial peer punishers should actively try to signal that they are calm and not driven by anger. This may be the most important direction of future research on peer punishment.

Conflict of Interest Statement

The Authors declare that there is no conflict of interest.

References

Balafoutas, L., & Nikiforakis, N. (2012). Norm enforcement in the city: a natural field experiment. *European Economic Review, 56,* 1773-1785. doi:10.1016/j.euroecorev.2012.09.008

Balliet, D., Mulder, L.B. & Van Lange, P.A.M. (2011). Reward, punishment, and cooperation: A meta-analysis. *Psychological Bulletin, 137,* 594-615. doi:10.1037/a0023489

Becker, G. (1968). Crime and punishment: An economic approach. *The Journal of Political Economy, 169,* 176-177.

Bell, M. D., Fiszdon, J. M., Greig, T. C., & Wexler, B. E. (2010). Social attribution test – multiple choice (SAT-MC) in schizophrenia: Comparison with community sample and relationship to neurocognitive, social cognitive and symptom measures. *Schizophrenia Research, 122,* 164-171. doi:10.1016/j.schres.2010.03.024

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. doi:10.1111/j.1559-1816.2005.tb02182.x

Bryant, F. B., & Smith, B. D. (2001). Refining the architecture of aggression: A measurement model for the Buss–Perry aggression questionnaire. *Journal of Research in Personality, 35,* 138–167. doi:10.1006/jrpe.2000.2302

Buhrmester, M. D., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data?. *Perspectives on Psychological Science, 6,* 3–5. doi: 10.1177/1745691610393980

Burnham, T. C. (2007). High-testosterone men reject low ultimatum game offers. *Proceedings of the Royal Society B: Biological Sciences, 274,* 2327–2330. doi:10.1098/rspb.2007.0546

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. doi: 10.1111/j.1559-1816.2008.00365.x

Cinyabuguma, M., Page, T., & Putterman L. (2006). Can second-order punishment deter perverse punishment? *Experimental Economics, 9,* 265-279. doi:10.1007/s10683-006-9127-z

Eriksson, K., Strimling, P., & Ehn, M. (2013). Ubiquity and efficiency of restrictions on informal punishment rights. *Journal of Evolutionary Psychology, 11,* 17-34. doi: 10.1556/JEP.11.2013.1.3

Eriksson, K., Cownden, D., Ehn, M., & Strimling, P. (2014). 'Altruistic' and 'antisocial ' punishers are one and the same. *Review of Behavioral Economics, 3,* 1–13. doi:10.1561/105.00000009

Gosling, S. D., Rentfrow, P. J., & Swann, W. B., Jr. (2003). A very brief measure of the big five personality domains. *Journal of Research in Personality, 37,* 504-528. doi:10.1016/S0092-6566(03)00046-1

Heider, F., & Simmel, M. (1944). An experimental study of apparent behavior. *American Journal of Psychology, 57,* 243-259.

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. doi:10.1006/jtbi.2000.2202

Ho, R., ForsterLee, L., ForsterLee, R., & Crofts, N. (2002). Justice versus vengeance: Motives underlying punitive judgements. *Personality and Individual Differences, 33,* 365-377. doi:10.1016/S0191-8869(01)00161-1

Horita, Y. (2010). Punishers may be chosen as providers but not as recipients. *Letters on Evolutionary Behavioral Science.* 1, 6—9.

Huesmann, L. R., & Guerra, N. G. (1997). Children's normative beliefs about aggression and aggressive behavior. *Journal of Personality and Social Psychology, 72,* 408-419. doi:10.1037/0022-3514.72.2.408

Kiyonari, T. and 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. doi:10.1037/a0011381

Klin, A. (2000). Attributing social meaning to ambiguous visual stimuli in higher-functioning autism and asperger syndrome: the Social Attribution Task. *Journal of Child Psychology and Psychiatry and Allied Disciplines, 41,* 831-846. doi: 10.1111/1469-7610.00671

Lagerspetz K., & Westman, M. (1980). Moral approval of aggressive acts. A preliminary investigation. *Aggressive Behavior,* 6, 119-130. doi:10.1002/1098-2337(1980)6:2<119::AID-AB2480060203>3.0.CO;2-Y

Levitt, S. D., & List, J. A. (2007). What do laboratory experiments measuring social preferences reveal about the real world?. *The Journal of Economic Perspectives,* 21, 153-174. doi: 10.1257/jep.21.2.153

Nikiforakis, N. (2008). Punishment and counter-punishment in public good games: Can we really govern ourselves?. *Journal of Public Economics, 92,* 91-112. doi:10.1016/j.jpubeco.2007.04.008

Nugier, A., Niedenthal, P. M., Brauer, M., & Chekroun, P. (2007). Moral and angry emotions provoked by informal social control. *Cognition and Emotion, 21,* 1699-1720. doi:10.1080/02699930601124738

Paolacci, G., Chandler, J., & Ipeirotis, P.G. (2010). Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making,* 5, 411–419.

Parks, C.D., & Stone, A.B. (2010) The desire to expel unselfish members from the group. *Journal of Personality and Social Psychology, 99,* 303-310. doi: 10.1037/a0018403

Pinker, S. (2011). Decline of violence: Taming the devil within us. *Nature, 478,* 309-311. doi:10.1038/478309a

Pratto, F., Sidanius, J., Stallworth, L. M., & Malle, B. F. (1994). Social dominance orientation: A personality variable predicting social and political attitudes. *Journal of Personality and Social Psychology, 67,* 741-763. doi: 10.1037/0022-3514.67.4.741

Ramirez, J. M. (1991). Similarities in attitudes toward interpersonal aggression in Finland, Poland, and Spain. *The Journal of social psychology, 131*(5), 737-739. doi: 10.1080/00224545.1991.9924656

Rimé, B., Boulanger, B., Laubin, P., Richir, M., & Stroobants, K. (1985). The perception of inter-personal emotions originated by patterns of movement. *Motivation and Emotion, 9,* 241-260. doi: 10.1007/BF00991830

Sterelny K. (2014). Cooperation, culture and conflict. *British Journal of Philosophy of Science.* Advance access, doi:10.1093/bjps/axu024.

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

Vandello, J. A., Ransom, S., Hettinger, V. E., & Askew, K. (2009). Men's misperceptions about the acceptability and attractiveness of aggression. *Journal of Experimental Social Psychology, 45,* 1209-1219. doi:10.1016/j.jesp.2009.08.006

Yamagishi, T. (1986). The provision of a sanctioning system as a public good. *Journal of Personality and Social Psychology, 51,* 110-116. doi: 10.1037/0022-3514.51.1.110

Yamagishi, T., Horita, Y., Mifune, N., Hashimoto, H., Li, Y., Shinada, M., ... & Simunovic, D. (2012). Rejection of unfair offers in the ultimatum game is no evidence of strong reciprocity. *Proceedings of the National Academy of Sciences, 109,* 20364-20368. doi:10.1073/pnas.1212126109

Table 1. Mean values per condition for measures in Study 1

Item	NO	CWP	IWP	CSP	ISP	Total	St Dev
Blue is generally trustworthy	1.87	1.46	1.33	0.90	0.73	1.26	0.84
Blue is generally angry*	1.77	0.96	0.65	0.38	0.24	0.80	0.91
Blue takes others' interests into account	1.66	1.60	1.33	1.22	0.78	1.32	0.85
Blue would punish others unfairly*	1.75	1.31	1.24	0.62	0.51	1.09	0.88
Blue generally follows standard norms of behavior	1.80	1.29	0.94	0.57	0.56	1.03	0.92
Blue's behavior was appropriate	6.19	4.70	4.51	3.06	2.77	4.24	2.14
A person behaving like Blue would be a problem for the group*	5.28	4.32	3.95	2.85	2.71	3.82	2.13
Would like to spend time with a person behaving like Blue	5.83	4.25	3.91	2.69	2.33	3.80	2.15
Blue approval index	0.79	0.24	0.03	-0.45	-0.62	0.00	0.80
Green's behavior was appropriate	6.23	5.51	5.55	4.91	5.88	5.62	1.49
A person behaving like Green would be a problem for the group*	5.24	4.97	4.95	4.46	4.84	4.89	2.10
Would like to spend time with a person behaving like Green	5.92	5.33	5.43	4.65	5.69	5.40	1.51
Pink's behavior was appropriate	6.21	5.49	5.62	4.76	5.88	5.59	1.50
A person behaving like Pink would be a problem for the group*	5.38	4.88	4.98	4.32	4.86	4.88	2.11
Would like to spend time with a person behaving like Pink	5.98	5.30	5.36	4.68	5.71	5.41	1.48
Green-Pink approval index	0.33	-0.04	0.00	-0.42	0.12	0.00	0.72

Note. Entries are mean values within each condition. The last two columns give the total mean and standard deviation. The scale is between 1 and 3 for the first five items, and between 1 and 7 for the remaining items (but not the indexes). Items marked * have been reverse-coded.

Table 2. Mean values of approval indexes per condition in Study 2

Measure	NO	CWP	IWP	CSP	ISP	Total	St Dev
Blue approval index	0.71	0.16	0.04	-0.37	-0.54	0.00	0.75
Green-Pink approval index	0.51	-0.04	-0.02	-0.54	0.11	0.00	0.82

Note. Entries are mean values within each condition. The last two columns give the total mean and standard deviation.

Table 3. Mean values per condition in Study 3

Item	NO	CWP	IWP	CSP	ISP	Total	St Dev
Blue's behavior was appropriate	5.51	5.58	5.35	4.79	4.78	5.20	1.60
A person behaving like Blue would be a problem for the group*	5.89	5.16	4.99	4.39	4.16	4.92	1.85
Would like to spend time with a person behaving like Blue	6.13	5.36	4.84	5.00	4.34	5.13	1.56
Blue approval index	0.45	0.16	-0.03	-0.20	-0.38	0.00	0.74
Green-Pink approval index	0.36	0.08	-0.08	-0.29	-0.08	0.00	0.71

Note. Entries are mean values within each condition. The last two columns give the total mean and standard deviation. The scale is between 1 and 7 for all items (but not the indexes). Items marked * have been reverse-coded.

Table 4. Mean values of Blue approval index per condition in Study 4

NO-B	IWP-B	ISE-B	ISE-un	IV	IVE	Total	St Dev
0.35	-0.62	-0.33	0.20	0.08	0.17	0.00	0.74

Note. Entries are mean values within each condition. The last two columns give the total mean and standard deviation.

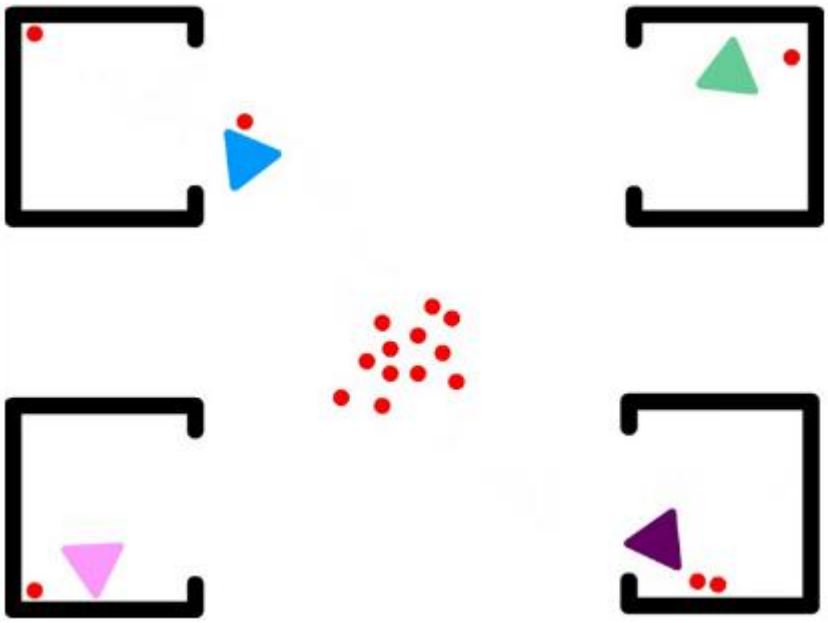

Figure 1. The triangles at their respective corners, with the blue triangle moving back from collecting a circle from the center.