**Third-party punishers do not compete to be chosen as partners**

Tommaso Batistoni1, Pat Barclay2 & Nichola J Raihani3

1. Centre for Experimental Social Sciences, University of Oxford, UK, OX1 1NF
2. Department of Psychology, University of Guelph, Canada, N1G 2W1
3. Department of Experimental Psychology, University College London, UK, WC1H 0AP

**Abstract**

Third-party punishment has been hypothesised to act as an honest signal of cooperative intent. Previous theoretical and empirical work has shown that individuals might escalate signals of cooperative intent when there is competition to be chosen as a partner. Here, we investigate the hypothesis that competition to be chosen as a social partner leads to escalating investment in third-party punishment. In the same scenario, we also consider the case of signalling via helpful acts to provide a direct test of the relative strength of the two types of signals. Investments in third-party helping were higher than investments in third-party punishment – and also exhibited a more robust positive association with audience effects. We did not find a clear effect of partner choice (over and above simply being observed) on either punishment or helping investments. Third-parties who invested more in helping were preferred as partners and were sent more money in a subsequent trust game. Third-party punishers were slightly preferred as interaction partners but less so than third-party helpers. In addition, we found that the amount invested in third-party punishment or helping was a reliable indicator of the individual’s trustworthiness: those who invested more returned a higher proportion of any entrusted amount. Individuals who did not invest in third-party helping were more likely to be untrustworthy, but the same was not true for individuals who did not invest in third-party punishment. This supports the conception of help as a less ambiguous signal of cooperative intent.

**Introduction**

Punishment is thought to be a key factor for maintaining cooperation among non-relatives (Fehr and Gächter, 2002, 2000; Raihani et al., 2012; Raihani and Bshary, 2019). Punishment refers to the act of paying a cost to inflict a reciprocal cost on a social partner (Clutton-Brock and Parker, 1995). As such, one has to ask how punishers might benefit from making these costly investments. One route to obtaining return benefits from punishment is if the target of punishment behaves more cooperatively in future interactions with the punisher (as originally suggested by (Clutton-Brock and Parker, 1995)). However, such outcomes seldom seem to occur in experimental settings (reviewed in (Raihani and Bshary, 2019)). Moreover, people often punish in situations where they were not the primary victim of the cheat and do not expect to interact with that individual in future interactions. Punishment in such settings has been termed ‘third-party punishment’ (Fehr and Fischbacher, 2004).

An alternative route by which punitive strategies could yield individual-level benefits to punishers is via reputation consequences that increase the punisher’s likelihood to have profitable social interactions in the future (Barclay, 2006; dos Santos et al., 2013, 2011; dos Santos and Wedekind, 2015; Jordan et al., 2016; Raihani and Bshary, 2015a, 2015b; Redhead et al., 2021). Building a reputation as a punisher might yield benefits in two distinct ways: (i) by signalling formidability, which can deter current social partners or bystanders from cheating when they interact with individuals with a punitive reputation (e.g. (dos Santos et al., 2013; Hilbe and Traulsen, 2012)); or (ii) by signalling cooperative intent, such that punishers benefit from increased access to cooperative interactions with new social partners (e.g. (Barclay, 2006; Dhaliwal et al., 2020; Jordan et al., 2016; Nelissen, 2008; Raihani and Bshary, 2015b)). Here we focus on the latter possibility.

Punitive acts can be conceptualized as signals that allow the punisher to convey an otherwise unobservable intent to cooperate (Jordan et al., 2016; Jordan and Rand, 2017; Przepiorka and Liebe, 2016). As long as the production of the signal (i.e. the punitive act) is associated with the hidden quality (i.e. cooperativeness), observers can then act contingently on the informative value of the signal and reward punishers for their actions (Raihani and Bshary, 2015b) or select punishers (over non-punishers) as partners (Jordan et al., 2016). If punishers stand to gain reputation benefits from punishing, then we might expect them to invest more in punishment when these decisions will be made known to other individuals (e.g. (Kurzban et al., 2007)). Nevertheless, the motives underpinning punishment decisions are hard to discern (Dhaliwal et al., 2020; Raihani and Bshary, 2019) meaning that the reputation consequences of punishing others might not always be positive (Horita, 2010; Ozono and Watabe, 2012; Przepiorka and Liebe, 2016; Raihani and Bshary, 2015a). In some cases, therefore, individuals might choose to hide investments in punishing others (Rockenbach and Milinski, 2011) or preferentially invest in helping rather than punishing others when these decisions will be revealed to others (Li et al., 2021).

Cooperative signalling becomes particularly salient when individuals are embedded in fluid social networks, and are therefore able to break and form social ties. The possibility to choose social partners can introduce a market-like logic in the realm of social interactions, resulting in an increased level of competition among individuals to be chosen by the best partners (Barclay, 2016, 2013; Noë and Hammerstein, 1995; Roberts, 1998). In this scenario, the ability to signal desirable qualities yields a strategic advantage as it allows signallers to attract high-quality social partners (or partners who are committed to cooperating, (Barclay and Barker, 2020; Roberts, 2020, Barclay et al. (forthcoming))) and, conversely, reduces the risk of interacting with exploitative partners. Confirming this perspective, previous works have shown that in so-called biological markets (Noë and Hammerstein, 1995)cooperation levels are higher than those observed compared to when no threat of social exclusion is present (e.g. (Barclay and Barker, 2020; Barclay and Raihani, 2016; Barclay and Willer, 2007); for recent reviews, see (Barclay, 2016; Hammerstein and Noë, 2016)).

Here, we investigate how punishers perform in such a biological market, under the hypothesis that punitive acts can function as costly signals of the punisher’s trustworthiness. Specifically, we ask whether individuals invest more in third-party punishment of a cheating individual when there is competition to be selected as a social partner by a bystander and, subsequently, whether punishers are preferred over non-punishers as social partners. We additionally assess the actual and perceived reliability of punitive acts as signals of trustworthiness by measuring whether investments in punishment are predictive of (i) the level of trust exhibited by a bystander when interacting with a punisher and (ii) the punisher’s trustworthiness in this interaction.

As previously argued (Raihani and Bshary, 2015a), the context in which punishment occurs is likely to determine the information it conveys and, therefore, its reputation consequences. Contexts where the punisher is harmed directly by the wrongdoer (i.e. second-party punishment) are more likely to be motivated by vengeful sentiments and are accordingly less likely to be interpreted as signals of the punisher’s cooperativeness. Conversely, when the punisher is an uninvolved bystander to the initial ‘crime’ (i.e. third-party punishment), then punitive acts are more likely to convey cooperative intent. As such, we explored the potential signalling value of punishment in a third-party punishment paradigm with partner choice, where there is a stronger theoretical argument for punishment to operate as a signal of cooperative intent. Punishment that is cheaper to administer than the damage it inflicts on a target could also be perceived as a competitive act (Raihani and Bshary, 2019, 2015a), rather than signalling an intent to cooperate. For this reason, in this study we used a fee-to-fine ratio of 1:1, meaning that punishers could not use punishment as a means to elevate their own payoffs relative to those of the target. Thus, we aimed to test the hypothesis that punishment might be used as a signal of cooperative intent under the most conducive conditions.

In addition to exploring whether individuals signal cooperative intent using punishment, we also aimed to replicate previous findings that costly investments in helpful behaviours also act as signals of cooperative intent and these investments increase when there is the possibility to be chosen by a partner for a subsequent social interaction (Barclay and Barker, 2020; Barclay and Willer, 2007; Sylwester and Roberts, 2010).

To summarize, in the current study we aim to extend the current signalling account of peer-punishment by testing whether:

* H1: the possibility for partner choice leads to higher investments in (a) third-party punishment and (b) costly helpful actions;
* H2: investments in (a) third-party punishment and (b) help positively predict probability to be chosen as a social partner by bystanders;
* H3: third-parties who invest more in (a) punishment and (b) help are trusted more by bystanders;
* H4: third-parties who invest more in (a) punishment and (b) help are more trustworthy.

**Methods**

**Participants.** All data, code and materials to reproduce this study can be found at { https://osf.io/4zpkb/}. This study was approved by the UCL Research Ethics Committee (project 3720/001). All data were collected in 2018 and participation was voluntary. We recruited 2253 participants through the online labour market Amazon Mechanical Turk (MTurk). Each participant was allocated to one of three roles: dictator (*n* = 902), third-party (*n* = 902) and bystander (*n* = 449). Throughout the study, roles were labelled using neutral terms. All participants received a show-up fee contingent on the role assigned ($0.20 for dictator, $0.50 for both third-party and bystander) and were given the chance to earn a bonus based on their decisions in the experiment. Total average earnings for each role were $0.52 (dictators), $1.12 (third-parties) and $0.80 (bystanders). All data were collected anonymously and no deception was used. The predictions in this study were not pre-registered.

**Experimental design.** Third-parties and bystanders were assigned to one of six treatments (described below). After reading the instructions, third-parties and bystanders completed a comprehension check (comprising 8 questions) and then made decisions in the experiment. Of the third-parties, 51.7 % participants answered all comprehension questions correctly; whereas 51.4 % of bystanders answered all comprehension questions correctly. All data are included in the main analysis to avoid selection bias, although we include comprehension as an explanatory term in our models. We also re-ran main models excluding participants who failed one or more comprehension checks and report qualitative differences.

The experimental setting consisted of three stages. In Stage A, dictators and third-parties played a variant of the Dictator Game (Kahneman et al., 1986). Each dictator was endowed with $0.50 and faced a dichotomous decision between a fair (“Keep $0.25 and give $0.25”) and an unfair (“Keep $0.45 and give $0.05”) share of the endowment with a passive receiver. Receivers were unrecruited MTurk workers who had taken part in a previous study run in our lab, who received a bonus according to the decision made by their matched dictator partner. Following the dictator decision, the third-party chose how much of their endowment, if any, to invest to punish an unfair Dictator or, according to the experimental condition, to help a receiver who was given an unfair share. Third-parties were endowed with $0.50 and could invest any amount (in $0.01 increments) between $0.00 and $0.45 to either help or punish the target individual, with a fee to fine/fee to help ratio of 1:1. These decisions were made using the strategy method and third-parties were informed that their decision would be implemented if they were matched to an unfair dictator.

In Stage B, third-parties were paired with another third-party, and one of each pair was selected (either randomly or actively chosen by a bystander, according to the experimental condition) to take part in Stage C.

In Stage C, the selected third-party and a bystander played a Trust Game (Berg et al., 1995) as the trustee and the trustor, respectively. The bystander was given $0.30 and could choose how much, if any, to send to the third-party. The amount sent was tripled and the third-party could choose what percentage (0%-100% in 10% increments) to keep for herself and what percentage to return to the bystander; the amount returned provides a measure of trustworthiness.

We implemented a 2 (Punish versus Help) x 3 (Random Allocation/No Knowledge; Random Allocation/Knowledge; Partner Choice/Knowledge) between-subjects fractional factorial design, resulting in six experimental treatments. Across treatments, we varied: *(i)* whether third-parties could punish an unfair dictator versus help the corresponding receiver in Stage A; *(ii)* whether third-parties were randomly allocated versus actively chosen by a bystander in Stage B and *(iii)* whether the bystander was informed versus not informed of third-parties’ behaviour in Stage A.

Allocations to all treatments occurred randomly within each session and participants made their decision in isolation. Participants were matched to a partner using ex-post matching (Rand, 2012). Across treatments, four third-parties could not be paired with another third-party and were therefore not matched with a bystander (but since decisions were collected using the strategy method, these are retained in the analysis). Details on the procedure and the matching protocol, as well as experimental instructions and comprehension questions are provided in the *Supplementary Information* (*SI*).

**Analysis**

All data were analysed using R (version 1.4.1106). This study was not pre-registered; hence we analysed our data using Bayesian and frequentist methods and report any qualitative differences in results. All analyses using frequentist methods are reported in the Supplementary Information. Below we report how we tackled each hypothesis, in turn. For all models, we fit cumulative link models to our data using brms (Bayesian regression models using Stan, (Bürkner, 2017; Carpenter et al., 2017)) assuming a probit link function. The Bayesian approach means that we estimate the full posterior distribution of parameters rather than single point estimates. We opt to present the Bayesian analyses in the paper as the posterior distribution is more informative than a single point estimate and more accurately accounts for uncertainty in modelling outcomes (Bürkner and Vuorre, 2019). We used cumulative link models because these allow dependent variables to be specified as ordinal categorical variables where the psychological distance between categories is not known; this is useful when response terms are not normally distributed (as was the case here) (Bürkner and Vuorre, 2019). Briefly, under this approach, the categorical response term is assumed to be drawn from an underlying latent continuous variable, *y*. The user can specify K thresholds in this response term, giving rise to K+1 categories. These thresholds are denoted by the model intercepts in the model output (and the number of intercepts is always K-1). The approach we use does not make any assumptions about the psychological distance between the thresholds; specifically, these are not assumed to be equidistant.

The output of models displays thresholds between the specified ordinal categories, rather than intercepts. Estimates associated with predictor variables are the posterior means of the parameters, which are analogous to frequentist point estimates. We also present 95 % credible intervals associated with these posterior means which specifies the interval within which the posterior mean falls with 95 % probability. All models have Rhat = 1 (indicating convergence) and effective sample size > 1000 (sufficient for stable estimates) unless stated otherwise. Our approach involved converting continuous response terms to ordinal categorical variables, which involves some subjectivity in determining the number of categories and where the thresholds between them should lie. These decisions were taken based upon visual inspection of the raw data, aiming to generate 3 categories for each dependent variable where the number of data points in each level was comparable. We provide this information below.

*H1. Does the possibility for partner choice lead to higher investments in (a) third-party punishment and (b) costly helpful actions?*

Based on visual inspection of the distribution of the data, we transformed punishment and help investments into ordinal categorical variables with 3 categories each (category 1 = $0.00, category 2 = $0.01-$0.20 and level 3 > $0.20). In the punishment condition, we had 255, 108 and 80 data points in categories 1, 2, and 3, respectively. In the help condition, we had 131, 182 and 146 data points in categories 1, 2 and 3, respectively.

To test whether third-parties invested more in punishment (or helping) when their decisions would be revealed to a future partner and/or when they might be chosen on the basis of their decision, each of these dependent variables was then fit to a separate CLM, with treatment (Anonymous / Random, Knowledge / Random, Knowledge / Choice) and comprehension (1/0, determining whether the participant correctly answered all the relevant comprehension questions) included as explanatory terms. We also included the two-way interaction between treatment and comprehension to account for the possibility that failed comprehension might impact investment more in some treatment conditions than in others. In all models, ‘Anonymous / Random’ was set as the reference category in treatment and we subsequently re-ordered the levels of this factor to test for differences between Knowledge/Random and Knowledge/Choice conditions.

*H2: Did investments in (a) third-party punishment and (b) help positively predict probability to be chosen as a social partner by bystanders?*

To address this hypothesis, we performed two binomial tests to determine whether the higher-investing punisher (or helper) was more likely to be chosen by the bystander for the subsequent Trust Game. We then used a Chi squared test to explore whether highest-investing helpers were more likely to be chosen than highest-investing punishers as partners.

*H3: Were third-parties who invested more in (a) punishment and (b) help trusted more by bystanders?*

We used a similar approach as for H1, by converting the amount entrusted to the third-party into a 3-level ordinal categorical variable (1 = sent $0.00, 2 = sent< $0.20, 3 = sent $0.20-$0.30) and setting this term as the dependent variable in a single CLM. We had 62, 98 and 138 data points in levels 1, 2 and 3, respectively. The following explanatory terms were included in the model: invest (the amount the third-party invested in punishment / helping), condition (1/0, denoting whether the condition was punishment or help) and the two-way interaction between these terms.

*H4: Were third-parties who invested more in (a) punishment and (b) help more trustworthy?*

We used a similar approach as for H1 and H3, by converting the proportion returned to the bystander into a 3-level ordinal categorical variable (1 = returned 0, 2 = returned < 40%, 3 = returned > 40 %) and setting this as the dependent variable in a CLM. We had 106, 116 and 227 data points in categories 1, 2 and 3 of the ordinal categorical response term. The same explanatory terms as for H3 were used.

**Results**

**H1: Does the possibility for partner choice leads to higher investments in (a) third-party punishment and (b) costly helpful actions?**

Most participants (255/443, 57.6%) did not punish the selfish dictator. The mean (± se) amount invested in punishment was $0.09 ± 0.01, with players investing anything from $0.00 to $0.45 to punish a third-party. Investing nothing was also the most common decision in the help condition, with 131/459 (28.5%) individuals investing nothing to help a receiver (range: $0.00-$0.45). However, the mean amount invested in help ($0.16 ± 0.01) was higher than mean amount invested in punishment (Wilcoxon test, W = 68998, p < 0.001).

Investment in third-party punishment was higher in both Knowledge / Random and Knowledge / Choice treatments, compared to baseline (Anonymous / Random), but these effects were not significant at conventional levels (95 % credible intervals for both posterior means included zero, Table 1, Figure 1a, Figure 2a). Due to high levels of failed comprehension in the task, we re-ran the model excluding the 128 / 443 participants who had failed one of the five comprehension questions concerning the incentives surrounding the investment in third-party punishment. In this reduced model, we detected a positive effect of Knowledge / Choice (relative to baseline) on punishment investment (estimate = 0.35, error = 0.17, CI: 0.02, 0.67). Analysing the same data using non-parametric frequentist methods, all effects were in the same direction, but people punished more in the Knowledge Choice treatment relative to baseline in the full sample but not in the reduced sample (full sample: p = 0.027, reduced sample: p = 0.16, full analyses presented in Supplementary Information). Investments in the Knowledge / Random condition were not significantly different to baseline (estimate = 0.24, error = 0.17, CI: -0.11, 0.56) and there was no significant increase in punishment investment in the Knowledge / Choice condition compare to Knowledge / Random (estimate = -0.12, error = 0.16, CI: -0.43, 0.21). Outputs based on frequentist analyses (see SI) were consistent with these latter two results.

Investment in third-party helping was affected by treatment, with third-parties investing more in the Knowledge / Choice treatment (compared to baseline, Table 2). When comparing the Random/Knowledge treatment to baseline, we find a discrepancy in results between the Bayesian and frequentist analyses: these two treatments were not significantly different in the Bayesian analysis (Table 2), but were significantly different in the frequentist analysis in both the full (p< 0.001) and reduced (p = 0.008) sample (see SI for all details). We found no difference between Knowledge / Random and Knowledge / Choice treatments (estimate = -0.29, error = 0.22, CI: -0.73, 0.14), a result that was also supported by our frequentist analyses.

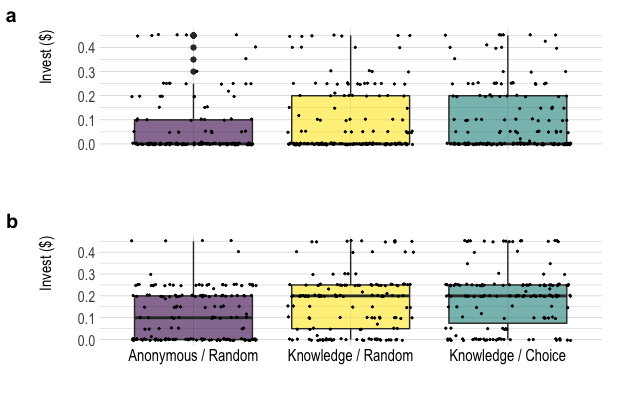
People who failed comprehension checks invested less in third-party helping (Table 2). Re-running the model, excluding the 164 / 459 participants who failed one of the five investment comprehension questions, revealed that investments were higher in both the Knowledge / Random treatment (estimate = 0.58, error = 0.16, CI: 0.25, 0.89) and the Knowledge / Choice treatment (estimate = 0.66, error = 0.16, CI: 0.35, 0.97), relative to baseline. Post-hoc comparisons revealed no difference between Knowledge / Choice and Knowledge / Random treatments (estimate = -0.08, error = 0.16, CI: -0.39, 0.24).

|  |  |  |  |
| --- | --- | --- | --- |
| **Parameter** | **Estimate** | **Error** | **Credible Interval** |
| Intercept 1 | 0.28 | 0.19 | (-0.08, 0.67) |
| Intercept 2 | 1.02 | 0.19 | (0.65, 1.40) |
| Treatment [Knowledge / Choice]\* | 0.40 | 0.26 | (-0.11, 0.92) |
| Treatment [Knowledge / Random] | 0.32 | 0.26 | (-0.17, 0.82) |
| Comprehension | -0.18 | 0.23 | (-0.62, 0.27) |
| Treatment [Knowledge / Choice]: Comprehension | -0.05 | 0.31 | (-0.66, 0.55) |
| Treatment [Knowledge / Random]: Comprehension | -0.07 | 0.31 | (-0.66, 0.54) |

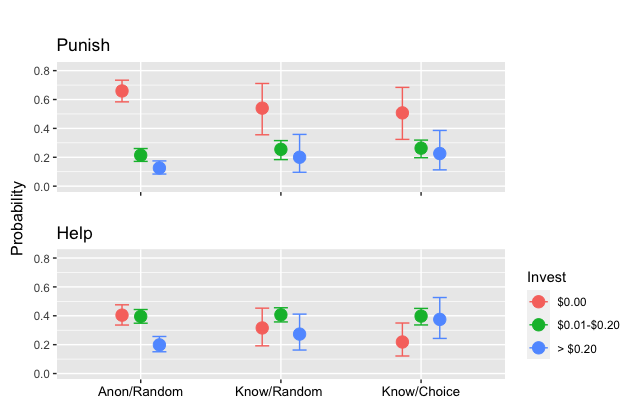
**Table 1. Investment in third-party punishment.** Intercepts represent the estimated thresholds between the 3 different categories (hence there are only 2 intercepts).These can be understood as the expected probability of a response falling in this category if all predictor variables were held at zero. Estimates are posterior means and error refers to posterior standard deviations, which are analogous to point estimates and standard errors, respectively. Here we display the model output including all participants and not the reduced model showing only participants who passed the comprehension check). \*This treatment is significant in the reduced model and in the frequentist analyses.

|  |  |  |  |
| --- | --- | --- | --- |
| **Parameter** | **Estimate** | **Error** | **Credible Interval** |
| Intercept 1 | -0.50 | 0.15 | (-0.78, -0.20) |
| Intercept 2 | 0.59 | 0.15 | (0.30, 0.88) |
| Treatment [Knowledge / Choice] | 0.53 | 0.21 | (0.12, 0.95) |
| Treatment [Knowledge / Random]\* | 0.24 | 0.21 | (-0.17, 0.66) |
| Comprehension | -0.40 | 0.18 | (-0.75, -0.04) |
| Treatment [Knowledge / Choice]: Comprehension | 0.14 | 0.26 | (-0.38, 0.64) |
| Treatment [Knowledge / Random]: Comprehension | 0.35 | 0.27 | (-0.16, 0.86) |

**Table 2. Investment in third-party helping.** Estimates are posterior means and error refers to posterior standard deviations. These are analogous to point estimates and standard errors, respectively. This model output includes all participants (and is not restricted to those who passed all comprehension checks). \* This treatment is significant in the reduced sample and in the frequentist analyses.



**Figure 1.** Boxplots of treatment on (a) third-party punishment and (b) third-party helping. Boxplots show the interquartile range and the median (in panel (a) the median =0 and is therefore not visible). Whiskers show min(max(x), Q\_3 + 1.5 \* IQR. Raw data points are overlaid with a jitter function. Larger black circles are outliers. Data are from the full sample.



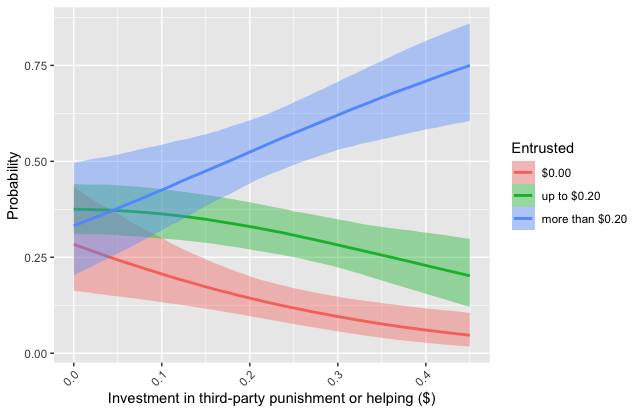
**Figure 2.** Posterior mean probabilities of individuals in each treatment being in each of the three investment categories according to whether they were punishing (top panel) or helping (bottom panel) a third-party. Error bars are 95 % credible intervals.

**H2: Did investments in (a) third-party punishment and (b) help positively predict probability to be chosen as a social partner by bystanders?**

In the punishment condition, 44 participants (22 pairs) pairs invested the same amount in punishment and could not be used in this analysis. In the remaining 56 pairs, we could ask whether the participant who invested the most in punishment was more likely to be chosen by the bystander as a partner. The highest investing player was chosen as the partner on 36 / 56 (64.3 %) occasions, indicating that bystanders used this information to select social partners (Binomial test, *p* = 0.04). In the help condition, 20 participants (10 pairs) invested the same amount in helping and could not be used in this analysis. In the remaining 65 pairs, the most helpful individual was chosen on 63 occasions (96.9 %, Binomial test, p< 0.001). The higher investor was more likely to be chosen in the help condition than in the punishment condition (Chi squared test, χ2 = 19.4, p< 0.001).

**H3: Were third-parties who invested more in (a) punishment and (b) help trusted more by bystanders?**

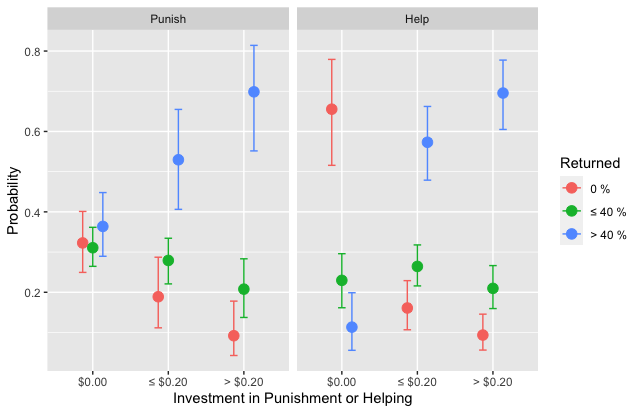
Bystanders entrusted on average $0.18 ± 0.01 (out of a possible $0.30) of their endowment to third-parties. Third-parties who invested more to help or punish were entrusted more by bystanders (estimate: 3.02, error = 0.94, CI: 1.18, 4.88, Figure 3) but bystanders did not entrust more to individuals that they chose to interact with compared to those they were randomly allocated to interact with (estimate: 0.15, error = 0.20, CI: -0.26, 0.54) or to helpers over punishers (estimate: -0.20, error = 0.23, CI: -0.67, 0.25). We found no evidence for two-way interactions between the amount the third party invested and either (i) being chosen by the bystander (estimate: -1.11, error = 0.97, CI: -3.00, 0.82) or (ii) whether the third party was a punisher or a helper on the amount entrusted by bystanders (estimate: 0.61, error = 1.06, CI: -1.49, 2.75; see also Table S3). There was a minimal effect of task comprehension on trust decisions, but credible intervals included zero (estimate =-0.03, error = 0.13, CI: -0.29, 0.24) and the results of this model do not qualitatively change when re-run excluding the 150 / 298 players who failed at least one comprehension question (Table S3).



**Figure 3. Amount entrusted as a function of investment in third-party help or punishment.** Plot shows posterior probability estimate for amounts entrusted to third-parties by bystanders. The plot indicates that the probability of sending higher amounts to the third-party increased with the third-party’s investment in punishment or helping. Conversely, the tendency to send nothing or a lower amount to the third-party had an inverse relationship with the third-party’s investment.

**H4: Were third-parties who invested more in (a) punishment and (b) help more trustworthy?**

Third-parties returned a mean of 33.4 (± 1.08) % of the endowment to bystanders. Third-parties who invested more in helping or punishing in Stage B were more trustworthy (estimate = 4.71, error = 0.66, CI: 3.41, 6.04) although we detected an interaction between investment and condition, whereby low investment in third-party help was a better predictor of untrustworthiness, compared to low investment in third-party punishment (estimate: -2.53, error = 0.91, CI: -4.32, -0.74; Figure 4). Most trustees (403 / 449) understood that they could maximise their payoffs by not sending any money back to the truster. Re-running this model excluding the 46 participants who failed that comprehension question did not qualitatively change results.



**Figure 4. How amount invested in punishment or helping impacted trustworthiness.** The plot shows predicted posterior means of increasing investment (specified as a categorical variable for ease of visualisation only) on return rates (specified as an ordinal categorical variable in the analysis). Panel 1 shows posterior probabilities in the help condition and panel 2 shows posterior probabilities in the punishment condition. Visual inspection reveals that low investment in helping predicts untrustworthiness (i.e. a high probability to return nothing to the truster), but the same is not true for low investment in punishment. For both help and punishment, increasing investments are associated with higher probability of trustworthiness.

**Discussion**

This study aimed to address a gap in the current literature by investigating whether investments in costly punishment are a signal of trustworthiness and whether such signals escalate in the presence of competition to be chosen as a partner. We found a clear effect of visibility on third-party helping decisions: individuals whose decision would be revealed to a future partner and/or who might be chosen on the basis of their investment tended to invest more in third-party help. However, we did not find that individuals invested significantly more when they could be chosen as a partner, compared to when partner choice was not possible – the effect was in the predicted direction but did not approach significance. The effects of observation and partner choice on third-party punishment were less clear: in our full sample, we detected no difference in punishment investment in either treatment, relative to baseline. When restricting our analysis to the subset of participants who correctly understood the incentives of the investment stage of the game, we found that participants invested more in third-party punishment when they could be chosen on the basis of this investment (compared to baseline). Our results also slightly differed depending on the analytic decisions made.

Investments in third-party punishment (and third-party helping) were positively associated with trustworthiness and also predicted the tendency to be selected for an interaction and entrusted with higher sums, which supports a signalling account of punishment (Barclay, 2006; Kurzban et al., 2007; Raihani and Bshary, 2015a, 2015b). However, whereas third-party helpers were overwhelmingly preferred as interaction partners, bystanders showed a weaker preference to interact with more punitive individuals. These results therefore support the conceptualisation of punishment as an ambiguous signal of cooperative intent (Raihani and Bshary, 2015a).

People who invested more in third-party punishment (and third-party helping) were more trustworthy on average, and were also more likely to be selected for an interaction and entrusted with higher sums, which supports previous results on punishment as a signal (Barclay, 2006; Kurzban et al., 2007; Jordan et al., 2016). However, whereas third-party helpers were overwhelmingly preferred as interaction partners, bystanders showed a weaker preference to interact with more punitive individuals. The stronger preference for third-party helpers may explain why our participants preferred to invest in helpful rather than punitive behaviour – a finding which replicates previous results (Li et al., 2021; Raihani and Bshary, 2015b). Overall, these results support the conceptualisation of punishment as an ambiguous signal of cooperative intent (Raihani and Bshary, 2015a).

Although many of our results align with previous results on helping and punishment, we note a failed replication of key results reported in Barclay and Willer (2007) and Sylwester & Roberts (2010). In those studies, the potential for partner choice caused an escalation of investments in costly helpful behaviour, relative to observation alone (see also (Barclay and Barker, 2020)). In our study, third-party investment did not significantly differ between conditions where behaviour was observable (but participants could not be chosen on the basis of their behaviour) and conditions where players could additionally be chosen on the basis of their investments. We note one discrepancy between our study and the previous ones that might explain this failed replication. In Barclay and Willer (2007) and Sylwester & Roberts (2010), participants chosen as partners subsequently played a Mutual Aid Game where they initially received a new endowment. Being chosen as partner for the second interaction, therefore, yielded direct access to new resources, as it did in some other studies showing evidence of competitive giving (Capraro et al., 2016, Giardini et al. in press) but see (Duffy and Kornienko, 2010). Conversely, in our design, chosen participants took part in a Trust Game as trustees and did not necessarily receive any resources from the new partner (because trusters could choose to send no money to the participant in the Trust Game). In this study, therefore, investments in costly signalling in the first interaction were imbued (perhaps more realistically) with a higher level of uncertainty regarding the benefits of being chosen as a partner. Similarly, any gains from the Trust Game might not have been enough to motivate participants to compete to help or punish, especially if they intended to be trustworthy and return half of what they were entrusted with. Another possibility is that the low comprehension rates limited the effect observed, despite our analyses which included only participants who correctly understood the incentive structure faced by third-parties.

We also note one unanticipated result, which was that players who invested nothing in third-party help were less trustworthy than players who invested nothing in third-party punishment. This result further supports the conceptualisation of help as a more reliable signal than punishment of future cooperative behaviour: investing nothing in help is a more reliable indicator that the player will be untrustworthy, whereas players who do not invest in punishment are not necessarily untrustworthy (Raihani and Bshary, 2015a). In our study, we tried to minimise the chance that punishment could be perceived as a competitive act: the punisher was not the victim of the original harmful act and the fee to fine ratio was set as 1:1 to rule out the possibility that individuals could elevate their own relative payoffs by enacting punishment (Raihani and Bshary, 2019). Nevertheless, even with these design features, punishment remains a fundamentally harmful act, involving accepting a cost to impose a cost on another individual. It seems apparent from this (and previous) work that whereas most cooperative individuals invest to help others, not all are willing to invest in third-party punishment. Understanding the roots of this variance (why do some cooperative individuals invest in punishment, while others do not?) is an important direction for future research.

We would like to note several limitations to this study. This study was conducted using an online platform and a predominantly Western sample – and, as such, the usual caveats on generalisability should be borne in mind. We also note that comprehension was relatively low and that our sample size was substantially reduced by excluding players who failed relevant comprehension checks. We also regret that our analyses were not pre-registered and, while our main hypotheses were generated a priori, several analytical decisions were taken after we had already collected the data and should be interpreted with this in mind.

Overall, our results support the idea that investments in third-party help (and, to a lesser extent, third-party punishment) signal trustworthiness and that observers take these actions into consideration when selecting and trusting partners. Our results are less clear on the idea that people compete with others to be chosen as interaction partners and we look forward to further work that will help to address this issue.

Acknowledgements:

Funding sources: PB is supported by the Social Sciences & Humanities Research Council of Canada (SSHRC grant #430287). NR is supported by the Royal Society and the Leverhulme Trust.

References

Barclay, P., 2016. Biological markets and the effects of partner choice on cooperation and friendship. Current Opinion in Psychology, Evolutionary psychology 7, 33–38. https://doi.org/10.1016/j.copsyc.2015.07.012

Barclay, P., 2013. Strategies for cooperation in biological markets, especially for humans. Evolution and Human Behavior 34, 164–175. https://doi.org/10.1016/j.evolhumbehav.2013.02.002

Barclay, P., 2006. Reputational benefits for altruistic punishment. Evolution and Human Behavior 27, 325–344. https://doi.org/10.1016/j.evolhumbehav.2006.01.003

Barclay, P., Barker, J.L., 2020. Greener Than Thou: People who protect the environment are more cooperative, compete to be environmental, and benefit from reputation. Journal of Environmental Psychology 72, 101441. https://doi.org/10.1016/j.jenvp.2020.101441

Barclay, P., Raihani, N.J., 2016. Partner choice versus punishment in human Prisoner’s Dilemmas. Evolution and Human Behavior 37, 263–271. https://doi.org/10.1016/j.evolhumbehav.2015.12.004

Barclay, P., Willer, R., 2007. Partner choice creates competitive altruism in humans. Proceedings of the Royal Society B: Biological Sciences 274, 749–753. https://doi.org/10.1098/rspb.2006.0209

Berg, J., Dickhaut, J., McCabe, K., 1995. Trust, Reciprocity, and Social History. Games and Economic Behavior 10, 122–142. https://doi.org/10.1006/game.1995.1027

Bürkner, P.-C., 2017. brms: An R Package for Bayesian Multilevel Models Using Stan. Journal of Statistical Software 80, 1–28. https://doi.org/10.18637/jss.v080.i01

Bürkner, P.-C., Vuorre, M., 2019. Ordinal Regression Models in Psychology: A Tutorial. Advances in Methods and Practices in Psychological Science 2, 77–101. https://doi.org/10.1177/2515245918823199

Capraro, V., Giardini, F., Vilone, D., Paolucci, M., 2016. Partner selection supported by opaque reputation promotes cooperative behavior. Judgment and Decision Making 11, 589–600.

Carpenter, B., Gelman, A., Hoffman, M.D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., Riddell, A., 2017. Stan: A Probabilistic Programming Language. Journal of Statistical Software 76, 1–32. https://doi.org/10.18637/jss.v076.i01

Clutton-Brock, T.H., Parker, G.A., 1995. Punishment in animal societies. Nature 373, 209–216. https://doi.org/10.1038/373209a0

Dhaliwal, N.A., Skarlicki, D.P., Hoegg, J., Daniels, M.A., 2020. Consequentialist Motives for Punishment Signal Trustworthiness. J Bus Ethics. https://doi.org/10.1007/s10551-020-04664-5

dos Santos, M., Rankin, D., Wedekind, C., 2011. The evolution of punishment through reputation. Proceedings of the Royal Society B: Biological Sciences 278, 371–377. https://doi.org/10.1098/rspb.2010.1275

dos Santos, M., Rankin, D.J., Wedekind, C., 2013. Human Cooperation Based on Punishment Reputation. Evolution 67, 2446–2450. https://doi.org/10.1111/evo.12108

dos Santos, M., Wedekind, C., 2015. Reputation based on punishment rather than generosity allows for evolution of cooperation in sizable groups. Evolution and Human Behavior 36, 59–64. https://doi.org/10.1016/j.evolhumbehav.2014.09.001

Duffy, J., Kornienko, T., 2010. Does competition affect giving? Journal of Economic Behavior & Organization 74, 82–103. https://doi.org/10.1016/j.jebo.2010.02.001

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

Fehr, E., Gächter, S., 2002. Altruistic punishment in humans. Nature 415, 137–140. https://doi.org/10.1038/415137a

Fehr, E., Gächter, S., 2000. Cooperation and Punishment in Public Goods Experiments. American Economic Review 90, 980–994. https://doi.org/10.1257/aer.90.4.980

Hammerstein, P., Noë, R., 2016. Biological trade and markets. Philosophical Transactions of the Royal Society B: Biological Sciences 371, 20150101. https://doi.org/10.1098/rstb.2015.0101

Hilbe, C., Traulsen, A., 2012. Emergence of responsible sanctions without second order free riders, antisocial punishment or spite. Scientific Reports 2, 458. https://doi.org/10.1038/srep00458

Horita, Y., 2010. Punishers May Be Chosen as Providers But Not as Recipients. Letters on Evolutionary Behavioral Science 1, 6–9. https://doi.org/10.5178/lebs.2010.2

Jordan, J.J., Hoffman, M., Bloom, P., Rand, D.G., 2016. Third-party punishment as a costly signal of trustworthiness. Nature 530, 473–476. https://doi.org/10.1038/nature16981

Jordan, J.J., Rand, D.G., 2017. Third-party punishment as a costly signal of high continuation probabilities in repeated games. Journal of Theoretical Biology 421, 189–202. https://doi.org/10.1016/j.jtbi.2017.04.004

Kahneman, D., Knetsch, J.L., Thaler, R., 1986. Fairness as a Constraint on Profit Seeking: Entitlements in the Market. The American Economic Review 76, 728–741. https://doi.org/10.2307/1806070?ref=search-gateway:ab3049daed7cb27ef2183218faa0c388

Kurzban, R., DeScioli, P., O’Brien, E., 2007. Audience effects on moralistic punishment. Evolution and Human Behavior 28, 75–84. https://doi.org/10.1016/j.evolhumbehav.2006.06.001

Li, Z., Hu, G., Xu, L., Li, Q., 2021. Third-Party Punishment or Compensation? It Depends on the Reputational Benefits. Front. Psychol. 12. https://doi.org/10.3389/fpsyg.2021.676064

Nelissen, R.M.A., 2008. The price you pay: cost-dependent reputation effects of altruistic punishment. Evolution and Human Behavior 29, 242–248. https://doi.org/10.1016/j.evolhumbehav.2008.01.001

Noë, R., Hammerstein, P., 1995. Biological markets. Trends in Ecology & Evolution 10, 336–339. https://doi.org/10.1016/S0169-5347(00)89123-5

Ozono, H., Watabe, M., 2012. Reputational benefit of punishment: Comparison among the punisher, rewarder, and non-sanctioner. Letters on Evolutionary Behavioral Science 3, 21–24. https://doi.org/10.5178/lebs.2012.22

Przepiorka, W., Liebe, U., 2016. Generosity is a sign of trustworthiness—the punishment of selfishness is not. Evolution and Human Behavior 37, 255–262. https://doi.org/10.1016/j.evolhumbehav.2015.12.003

Raihani, N.J., Bshary, R., 2019. Punishment: one tool, many uses. Evolutionary Human Sciences 1. https://doi.org/10.1017/ehs.2019.12

Raihani, N.J., Bshary, R., 2015a. The reputation of punishers. Trends in Ecology & Evolution 30, 98–103. https://doi.org/10.1016/j.tree.2014.12.003

Raihani, N.J., Bshary, R., 2015b. Third-party punishers are rewarded, but third-party helpers even more so. Evolution 69, 993–1003. https://doi.org/10.1111/evo.12637

Raihani, N.J., Thornton, A., Bshary, R., 2012. Punishment and cooperation in nature. Trends in Ecology & Evolution 27, 288–295. https://doi.org/10.1016/j.tree.2011.12.004

Rand, D.G., 2012. The promise of Mechanical Turk: How online labor markets can help theorists run behavioral experiments. Journal of Theoretical Biology 299, 172–179. https://doi.org/10.1016/j.jtbi.2011.03.004

Redhead, D., Dhaliwal, N., Cheng, J.T., 2021. Taking charge and stepping in: Individuals who punish are rewarded with prestige and dominance. Social and Personality Psychology Compass 15, e12581. https://doi.org/10.1111/spc3.12581

Roberts, G., 2020. Honest signaling of cooperative intentions. Behav Ecol. https://doi.org/10.1093/beheco/araa035

Roberts, G., 1998. Competitive altruism: from reciprocity to the handicap principle. Proceedings of the Royal Society of London. Series B: Biological Sciences 265, 427–431. https://doi.org/10.1098/rspb.1998.0312

Rockenbach, B., Milinski, M., 2011. To qualify as a social partner, humans hide severe punishment, although their observed cooperativeness is decisive. PNAS 108, 18307–18312. https://doi.org/10.1073/pnas.1108996108

Sylwester, K., Roberts, G., 2010. Cooperators benefit through reputation-based partner choice in economic games. Biology Letters 6, 659–662. https://doi.org/10.1098/rsbl.2010.0209