When Do Punishment Institutions Work?

Patrick Aquino[†]
Robert S. Gazzale[‡]
Sarah Jacobson[§]

August 2015

Abstract

While peer punishment sometimes motivates increased cooperation, it sometimes reduces cooperation. We use a lab experiment to study why punishment sometimes fails. We begin with a gift exchange game with punishment as it has typically been implemented therein since punishment has often backfired in this game. We modify two features of punishment that could increase its efficacy: punishment's strength and its timing (whether the punisher publicly pre-commits to punishment or acts after the punishee). We replicate the result that peer punishment in gift exchange games can reduce cooperation, but show that this bad outcome disappears if punishment is more powerful. This does not seem primarily due to punishment's threat leading to spiteful behavior: we find little evidence of spite, and the same punishment does not perform better when it is chosen after the fact. We find two main reasons that punishment decreases cooperation: lower wages are offered (a stick is substituted for a carrot); and many punishers don't design punishment to properly incentivize high effort, particularly when punishment is weak in power. Punishment that is not publicly pre-committed is not effective in this game, even though this kind of punishment is similar to that used in public good games in the literature where punishment does seem to increase cooperation. The only punishment institution that increases cooperation is high-power punishment that is publicly pre-committed, which works through strong incentives rather than reciprocity. Finally, the existence of a punishment institution often decreases social surplus (when punishment-related losses are considered), although it may eventually increase social surplus if it is powerful and publicly pre-committed.

JEL Classifications: D03, C91, D64, J49, H41

Keywords: punishment, cooperation, gift exchange, reciprocity

[†] Deerfield Academy, Deerfield, MA; pa.aquino@gmail.com.

[‡] Department of Economics, University of Toronto, 150 St. George Street, Toronto, ON M6G 1W2; robert.gazzale@utoronto.ca.

[§] Corresponding author. Department of Economics, Williams College, 24 Hopkins Hall Dr., Williamstown, MA 01267; sarah.a.jacobson@williams.edu.

1 Introduction

In important situations ranging from employer-employee relations to management of a communal fishing area and beyond, good behavior is voluntary but is social-surplus-increasing. Societies and firms must establish institutions to encourage cooperation in such settings. Peer punishment is easy to implement and sometimes increases cooperation, but in other cases it does not, and sometimes it even causes worse outcomes. Under what circumstances does peer punishment fail?

The purpose of this paper is to use laboratory experiments to replicate the "backfiring" of peer punishment observed in some studies and to explain what features of the punishment institution cause it. We use a gift exchange game (first used to model employment relations in Akerlof, 1982, and initially studied in the lab in Fehr et al., 1993). We focus on two features of the punishment institution in this game. First, we vary whether there is pre-commitment to determine whether pre-committed punishment is *per se* detrimental as others have suggested. Second, we vary the strength (i.e., size) of punishment. In studies in which punishment backfires, the punishment that the punisher can levy is generally weak or small (as in Fehr et al., 1997). By comparison, punishment that reinforces cooperation in public good games tends to be relatively powerful (e.g. one token spent by the punisher reduces the punishee's payoff by three tokens, as in Fehr and Gächter, 2002a). We conjecture that this weak power of punishment is part of the reason punishment has failed (many studies have shown that punishment becomes more effective as it is stronger, as discussed in Putterman, 2014), and further that power and timing of punishment may interact with each other.

As summarized in Putterman (2014), peer punishment in social dilemmas has been studied for decades. Seminal studies in this area include Ostrom et al. (1992) and Fehr and Gächter (2000), who find that adding a punishment institution to a public good game increases cooperation and prevents the decline in contributions that typically occurs in such games. However other studies find that the result does not extend to other institutions. In Fehr et al. (1997), subjects in a gift exchange game with peer punishment actually see lower cooperation levels as compared to the no-punishment baseline. Other studies (e.g., Fehr and Gächter, 2002b, Fehr and Rockenbach, 2003) have found similar results. Some in this literature argue that incentive-based punishment—punishment that self-interested punishers would willingly mete out and that would change the best response action of money-maximizing punishees—is unsuccessful because it crowds out trust or creates bad emotions, while reciprocity-based punishment—which punishers use to reciprocate bad acts but which is generally not money-maximizing for them to choose—is successful because it is deemed more legitimate and leverages social preferences (see Fehr and Falk, 2002).[1] Indeed, the punishment that "backfires" in Fehr et al. (1997) is publicly pre-committed and is an equilibrium choice of self-interested actors, while in that study and other studies of gift exchange games (and trust games) punishment that is not publicly-pre-committed is often associated with better outcomes (e.g., Fehr and Gächter, 1998, Calabuig et al., 2013, Rigdon, 2009; also, punishment that

[1] The failure of pre-committed punishment might also be due to the *substitution heuristic* (Kahneman and Frederick, 2002). When faced with the difficult question "Which action should I take?", the agent makes his choice based on the easier-to-answer question "Which action results in the highest monetary payoff?" when his true preferences may include social factors as well.

increases public good game cooperation is usually decided on after the fact as in Fehr and Gächter, 2000). However, the design of punishment institutions in these games has varied, so it has been difficult to identify precisely when peer punishment will fail and why.

This paper contributes to the literature by directly identifying factors that cause the existence of a peer punishment institution to reduce cooperation.

We find that publicly-pre-committed punishment that is weak in power, as it typically has been in games in which the backfire has been observed, does reduce cooperation relative to the no-punishment baseline. However, it's not quite true that this is caused by crowding out of positive feelings. We find little evidence of spiteful underperformance, and punishment that cannot be publicly pre-committed is ineffective in increasing cooperation even when it is relatively strong. One of the major reasons that peer punishment backfires in that scenario, we show, is that when punishment is available, the punisher shows less trust (sends a lower wage): they substitute from the "carrot" toward the "stick," and since trust is reciprocated with cooperation, cooperation is reduced. We also show that punishers frequently do not choose punishments designed to elicit high cooperation, and this is particularly the case when punishment is weak. The punishment institution we study that is most effective is pre-committed and strong—so it relies on incentives but provides strong incentives, and punishers seem to learn to structure these incentives properly as they play repeated rounds. Finally, we echo the result from the literature (see Putterman, 2014) that the costs incurred by punishment often outweigh the social benefits of increased cooperation, except at the the very end of our treatment with strong pre-committed punishment.

2 Experiment Design

We first present a broad overview of our experiment design. We then provide details.

2.1 Overview

The design is across-subject: each subject participates in only one treatment. All subjects in a session experience the same treatment. Each session consists of 10 rounds. Each round is a one-shot interaction between a principal and an agent based on the gift-exchange experiments in Fehr et al. (1997). The experiment follows a double-blind (or double-anonymous) procedure (e.g., Hoffman et al., 1996): subjects do not know which other subjects they are interacting with, and subject choices cannot be matched to a particular person even by the experimenters.

In each round, a principal specifies a wage: the number of points to be paid from the principal to the agent at the end of the round. After learning the wage, the agent chooses one of four effort levels. Higher effort levels result in higher payments to the principal but are more costly to the agent. In treatment sessions, principals can punish agents. Except as noted, we use the strategy method to elicit the principal's complete punishment profile.[2] Before learning the agent's choice, the principal specifies for each effort level the number of punishment tokens she will purchase if that

[2] Casari and Cason (2009) show that the strategy method can lead to lower trustworthiness; this is not what we observe in our data.

effort level is chosen, with agent point reductions proportional to the number of tokens purchased.

We use a full factorial design, varying design along two dimensions. In one dimension, we vary when the principal chooses punishment. In the Ex Ante treatments, the principal chooses a punishment profile that is shown to the agent *before* he chooses an effort level. In this case, the principal publicly pre-commits to punishment, and is therefore the first mover in a two-stage game. In the Ex Post treatments, the principal chooses punishment *after* the agent chooses an effort level (in one case by direct elicitation, i.e. in response to the agent's choice of effort; in other cases by strategy method elicitation, i.e. without seeing the agent's effort choice), and the agent learns the punishment chosen by the principal *after* the agent chooses an effort level. In the second dimension, we vary the strength of the punishment. In all treatments, the principal chooses from zero to five punishment tokens for each possible agent action (or for the agent's actual action, in the direct elicitation case), and each token the principal purchases costs her one point. (The principal only purchases and pays for the tokens specified for the action actually chosen by the agent.) In Weak treatments, each punishment token actually purchased reduces the agent's payoff by one point, whereas in Strong treatments, each punishment token actually purchased reduces the agent's payoff by three points.

2.2 Details

For ease of exposition, we use an employer-employee context to explain our experiment design. In the actual experiment and instructions, the context is neutral: we refer to the principal as "Role 1" and the agent as "Role 2;" wage as "transfer;" effort as "action;" and punishment as "reduction."

At the start of a session, each of 18 subjects is randomly assigned to a computer, implicitly assigning her to one of three equally-sized groups. Each subject receives a written copy of the instructions (see Appendix A), which are read aloud. Before decision-making rounds, subjects answer review questions to ensure they understand the procedures and how payments are calculated. Only when all subjects have correctly answered all review questions do the rounds commence. While making decisions, each subject can see a history box containing all information previously revealed to the subject in this and preceding rounds. At the end of each session, subjects complete a brief demographic questionnaire. One of the ten rounds is randomly selected to determine payment. Experiment earnings are $1 for every four points earned in the selected round. Subjects receive their payments completely privately to maintain anonymity even from the experimenters.

Throughout the session, a subject only interacts with the five other subjects in her group. A subject anonymously interacts with each member of her group twice: once as principal and once as agent. This means that each subject is principal in five rounds and agent in five rounds.[3] However, subjects do not know when they are interacting with each group member, so individual reputation and between-round reciprocation cannot affect their choices. Within these constraints, and the added constraint that each subject must switch roles between the first and second round,

[3] The fact that subjects play both roles could make them more sympathetic to the other role. This could reduce the use of punishment as compared to a game that does not use role reversal, and with less punishment could come less cooperation. However, this sympathy could also increase cooperation directly.

the subject's role and partner are randomly determined in each round.

Table 1: Payoff consequences of effort levels

Effort Level	0	1	2	3
Cost to Agent	0	4	10	18
Benefit to Principal	0	30	60	90

To ensure non-negative and relatively equal earnings, the principal starts each round with 100 points and the agent with 20. At the start of a round, the principal chooses how many points to transfer to the agent. This wage can be any integer from 20 to 90 and is not contingent on the agent's effort. Upon learning the transfer, the agent chooses one of four effort levels. As shown in Table 1, effort has an increasing marginal cost to the agent but yields a much higher constant marginal benefit of 30 to the principal.

In treatment sessions, except for one treatment as discussed below, the principal specifies a punishment profile. That is, for each of the four possible effort levels, the principal indicates the number of punishment tokens she will purchase if the agent chooses that effort level. She can choose any integer between and including zero and five tokens. The principal only pays for the punishment tokens she specifies for the action actually chosen by the agent. The principal pays 1 point for each token actually purchased. In Weak treatments, each punishment token actually purchased reduces the agent's payoff by one point, whereas in Strong treatments, each token reduces the agent's payoff by three points.

Finally, the distinction between Ex Ante and Ex Post sessions is that in Ex Ante, the principal chooses punishment and the agent observes that choice before choosing his effort level, whereas in Ex Post the principal chooses punishment after the agent chooses effort and the agent therefore does not know what punishment he faces until after he chooses effort. In other words, in Ex Ante, the principal publicly pre-commits to punishment, whereas in Ex Post she does not. We use two versions of Ex Post. In Ex Post Direct, the principal chooses a single punishment level after learning the agent's effort choice. In Ex Post Strategy, the principal does not yet observe the agent's effort choice and chooses an entire punishment profile. After the principal makes her punishment choice, she learns the agent's effort choice and the agent learns the principal's punishment decision. Therefore, between Ex Ante and Ex Post Strategy there is no difference in the information provided to the agent over the course of the round, although there is a difference in timing. The difference in timing can also be thought of as changing the information set available to the agent (whether he knows what punishment he faces) when the agent is choosing effort.

To put agents in the Ex Post treatments on more even footing with agents in Ex Ante sessions with regard to expectations about punishment, principals in Ex Post request a desired effort level from their agents. This request is cheap talk, as the experiment does not force agents to fulfill that effort level and principals need not (and often cannot) specify a punishment profile that would make that effort level payoff-maximizing, much less force agents to choose that effort.

3 Theory

In this section, we derive theoretical predictions for the principal-agent game used in our experiment. We start by establishing the Subgame Perfect Nash Equilibria (SPNE) of the one-shot interaction for the treatments under the assumption that subject preferences depend only on subject earnings. We then discuss how these predictions change if subjects are also driven by positive or negative reciprocity. Although the experiment involves 10 one-shot rounds, since subjects are rematched with new partners for each round, each round is independent and thus we only examine predictions for the one-shot game.

The SPNE of the Baseline (no-punishment) principal-agent game is straightforward to find through backward induction. The agent always chooses 0 effort, and therefore the principal always pays the minimal wage (20). We have the same equilibrium prediction in the Ex Post Direct treatment. The principal never chooses costly punishment. Because the threat of punishment is not credible, the agent considers only the direct cost of effort and chooses 0 effort, and thus the principal always pays the minimal wage (20).

Whereas no effort is predicted in the Baseline and Ex Post Direct conditions, the equilibrium in each Ex Ante treatment has positive effort. Recall that the principal can buy up to 5 punishment tokens at a cost of 1 point per token, and each token reduces the agent's payoff by 1 point in Weak and 3 points in Strong treatments. From Table 1, we see the cost of effort is 0 for the lowest effort level (0) and is 4 for an effort of 1, and that each unit of effort earns the principal 30 points. As the principal can reduce the agent's payoff from zero effort by more than 4 points, the principal can and will induce the agent to choose effort of 1 whether punishment is Weak or Strong. The agent's cost for an effort of 2 is 10. Only in the Strong treatments can the principal increase the cost of choosing effort of 0 by 10 or more points, and thus only in Strong Ex Ante can the principal induce the agent to choose effort of 2. In neither Ex Ante treatment can the principal reduce the payoff to effort of 0 by 18 points, so in neither treatment can the principal induce the agent to choose effort of 3 (which costs the agent 18 points). Of course, because effort is still not predicted to be affected by wage, the equilibrium wage is still the minimal wage of 20.

In Ex Post Strategy, because a principal does not see the agent's effort before she chooses a punishment profile, it is arguably as if she chooses the punishment profile at the same time as the agent chooses an effort level. Thus, the Ex Post Strategy game can be conceived of as a simultaneous-move game in which the only subgame is the game itself. The principal is best responding as long as her profile has no punishment for the agent's chosen effort. One equilibrium is that characterized in Baseline and Ex Post Direct: the agent chooses zero effort and the principal does not punish for zero effort. It also includes the equilibria from Ex Ante because these equilibria result in no punishment, and many other possible move combinations. In these equilibria, however, the principal suffers a cost to punish the agent if the agent does not choose an effort that results in zero punishment. We argue that since the principal cannot credibly commit to this costly punishment, equilibria with non-zero punishment (and non-zero effort) are less plausible than the no effort, no punishment equilibrium. If this is so, behavior in the Strong Ex Post Direct and

Strategy treatments should be the same. As we discuss later, behavior does differ, but not because of different punishment—rather, because of different wages and reciprocity.

In summary, in all cases, the wage is the minimal wage. In the Baseline and the Ex Post Direct treatment, equilibrium effort is 0, with no punishment in Ex Post Direct. No effort, and no punishment for this choice, is the most plausible equilibrium for Ex Post Strategy. In Weak Ex Ante, equilibrium effort is 1, which is supported the principal choosing a punishment profile of (≥ 4, 0, 0, 0) (that is, 4 or 5 punishment tokens if the agent chooses 0, and 0 if the agent chooses 1, 2, or 3). In Strong Ex Ante, equilibrium effort is 2, which is supported the principal choosing any profile that satisfies (≥ 4, ≥ 2, 0, 0).[4,5]

We now qualitatively discuss how the analysis changes if agents have preferences for anything other than monetary consequences. In particular, we consider reciprocal preferences—agents willing to pay a monetary costs to reward those who have been "nice" and to punish those who have not, i.e., both positive and negative reciprocity per the terminology of Cox and Deck (2005).

In all of our treatments, wage offered can be interpreted as trust and the responsivenes of effort to wage can be interpreted as reciprocity. Fehr et al. (1997) and Fehr and Gächter (2002b) find evidence of agents reciprocating generous wage offers from the principal. We thus expect at least some principals to extend generous wage offers, with the extent of the generosity likely depending on the extent to which they are reciprocated. Note that the principal need not have other-regarding preferences to offer a wage; she need merely expect the agent to reciprocate. If the principal does have other-regarding preferences, e.g. altruism, she may offer a wage regardless of her expectations.

Reciprocal preferences may play an additional role in the Ex Ante treatments. As in Fehr et al. (1997), agents may view committed-to punishment as displaying a lack of trust, and may respond negatively. Intrinsic motivation to cooperate may be crowded out by this extrinsic motivation (as in Gneezy and Rustichini, 2000), in which case agents will reduce their voluntary cooperation level (i.e., stop providing as much effort above their payoff-maximizing level). However, agents may respond by choosing effort levels so low that they suffer punishment to avoid increasing the payoff of the principal. This is spite, i.e., negative reciprocity (Cox and Deck, 2005). Since principals in this treatment can choose to offer both a carrot (a generous wage) and a stick (punishment for low effort), it is not *a priori* obvious how these two tools will interact in agents' minds as a signal of trust, or how sophisticated principals may be in their guesses about agents' behavioral reactions.

In the case of Ex Post punishment, negative reciprocity or norm enforcement may inspire a principal to punish after the agent has acted. This intuition is consistent with findings of Fehr and Gächter (2000) and others. If agents are sophisticated, they should expect this punishment, i.e., the threat to punish should be credible. An agent who does not reciprocate a generous wage offer might be seen as unkind, and thus we might expect that the principals most willing to punish will

[4] At equilibrium, agents are never actually punished. This means that the SPNE we identify for the Ex Ante treatments are also Nash equilibria (although not SPNE, i.e., they are off-equilibrium-path outcomes) of the Ex Post treatments. Relatedly, there are SPNE that differ from our main focal SPNEs in punishment specified at off-equilibrium effort levels: for example, in Weak Ex Ante, any strategy that meets (≥ 4, 0, ≥ 0, ≥ 0) is an equilibrium.

[5] These punishment profiles weakly support the efforts of 1 in Weak and 2 in Strong Ex Ante: the agent is indifferent between effort of 0 and 1 if punishment at effort of 0 is 4, and between effort of 1 and 2 if punishment at effort of 1 is 2. A strict equilibrium requires punishment profiles of (5, 0, 0, 0) in Weak and (≥ 4, ≥ 3, 0, 0) in Strong Ex Ante.

be those who have paid generous wages. Agents who are fully self-regarding may be disciplined into giving higher effort if punishment is frequent and severe enough. Agents with other-regarding preferences may particularly wish to conform to norms of cooperation signaled by punishment, and thus may cooperate more than is supported by the punishment they expect.

4 Results

The experiment was run in January 2011 and February 2015 at Williams College. Two hundred and sixteen undergraduate students participated as subjects in twelve sessions (2 sessions each of the Baseline and of each treatment), with eighteen subjects in each session. Therefore, there are thirty-six subjects per treatment, each of whom was a principal for five rounds and an agent for five rounds.[6] Subjects were recruited through the online recruitment system ORSEE (Greiner, 2004). The experiment was programmed in z-Tree (Fischbacher, 2007). Sessions lasted approximately 60–75 minutes. Subjects earned \$17.80 on average. The subjects' average age was 19.79, and 52.78% of subjects were female.

In the analysis that follows, we use individual subjects as the unit of observation since rematching of subjects between rounds renders individuals within a session effectively independent. We will conduct within- and between-subjects univariate nonparametric tests at the individual level and panel regression analysis at the individual-round level.

4.1 A Peer Punishment Institution Can Reduce Cooperation

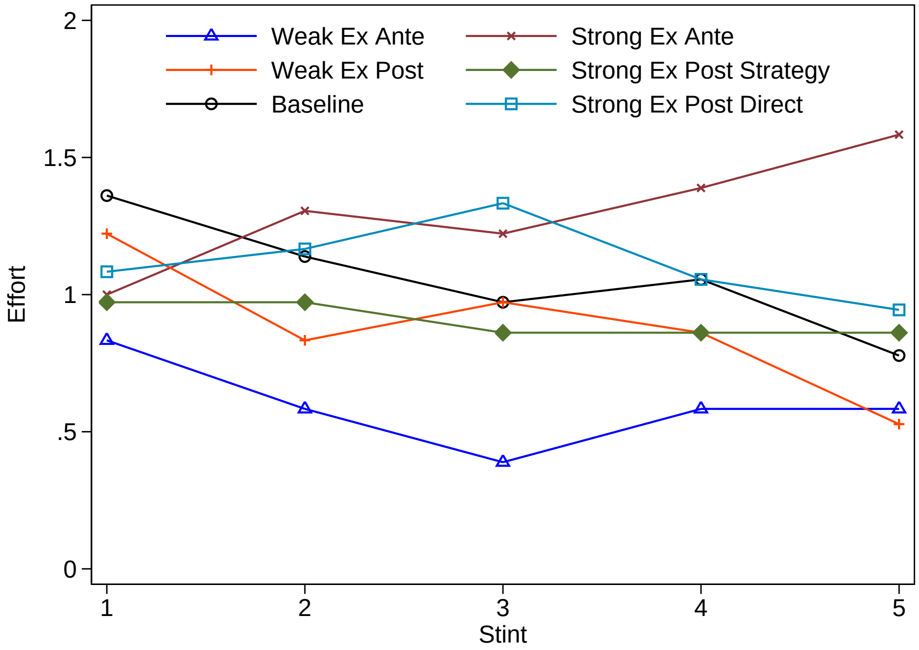
We replicate the result from Fehr et al. (1997) and others that punishment as it has typically been implemented in gift exchange reduces cooperation. Figure 1 shows the average effort of agents across periods.[7] Weak Ex Ante clearly provides lower effort than Baseline. Strong Ex Ante performs better, particularly as the stints progress, and the Ex Post treatments look similar to Baseline.

Table 2 shows average effort across the treatments. Weak Ex Ante's significantly lower effort as compared to Baseline replicates the backfire of peer punishment observed in Fehr et al. (1997).

Our results show that it is not exactly the case that peer punishment backfires because the timing of ex ante punishment causes bad feelings. Effort in Ex Post is better than Ex Ante for Weak, but we show in Section 4.2 that that is explained by the higher wage offered. For Strong, Ex Post actually performs worse than Ex Ante. As we will show in Section 4.3, punishment is not used nearly as much in Strong Ex Post as in Strong Ex Ante, presumably because it is not as credible of a threat. Therefore, the punishment in Strong Ex Post does not have the same deterrent effect as in Strong Ex Ante. The Direct treatment does yield more effort than the Strategy treatments ($p = 0.056$ when compared to Weak, $p = 0.066$ when compared to Strong), but is still statistically no greater than Baseline. Since the Strong Ex Post treatments use punishment more like that used in public good games like (Fehr and Gächter, 2000), it is interesting that neither provides higher

[6] Effort, wage, and punishment did not vary in any treatment by whether a subject was agent or principal first.

[7] Since different subjects are agents in different periods, Figure 1 and subsequent figures plot average values for each subject's first, second, third, fourth, and fifth stint in the role of agent (or principal, as appropriate). The trend is similar if we instead plot behavior across the 10 rounds.



Vertical axis is effort, which can be 0, 1, 2, or 3. Horizontal axis is stint in role.

Figure 1: Average Effort by Stint Across Treatments

Table 2: Differences in Effort Across Treatments

Treatment	Effort	Diff vs Baseline?	Weak & Strong Diff?	Ex Ante & Ex Post Diff?
Baseline	1.06 (0.71)			
Weak Ex Ante	0.59 (0.61)	0.006	0.000	0.039
Strong Ex Ante	1.30 (0.68)	0.188		0.013
Weak Ex Post	0.88 (0.66)	0.253	0.821	
Strong Ex Post Strategy	0.91 (0.61)	0.311		
Strong Ex Post Direct	1.12 (0.46)	0.852		
N	36 each			

Standard deviations in parentheses. "Diff" columns present p-values of Wilcoxon rank-sum tests across treatments.

effort than Baseline. We will speculate later on this difference relative to the literature.

Further, Table 2 shows that the detrimental effects of punishment in Weak Ex Ante disappear when the strength of punishment is increased. We show in Section 4.3 that this is because some of Weak Ex Ante's poor performance is caused by poorly-chosen punishment profiles; principals tend to design better profiles when punishment is Strong. In fact, as shown in Figure 1, effort increases across stints in Strong Ex Ante. If we exclude subjects' first stint, the Strong Ex Ante treatment shows effort significantly greater than Baseline (Wilcoxon rank-sum $p = 0.034$). We show in Section 4.3 that this is likely because principals learn to specify profiles that incentivize high effort.

4.2 The Role of Wage

One reason peer punishment backfires is because principals reduce the wage they offer when they have a plausible punishment tool.

First, we demonstrate this by performing OLS regressions of effort on treatment dummies and principal choices, as shown in Table 3.[8, 9] We will focus on wage-related results here, and we will discuss the regression results related to punishment in Section 4.3.

Specification (1) reprises the result from Table 2 that Weak Ex Ante performs worse than the no-punishment Baseline. But Specification (2) shows that wage is significantly reciprocated and that for a given wage, effort in Weak Ex Ante is the same as in the Baseline. Indeed, although it is not highlighted as a potential cause of reduced effort, Fehr et al. (1997) see lower wages when punishment is possible. Table 2 also shows that for a given wage, effort is higher in Strong Ex Ante than in Baseline. We will show in Section 4.3 that this is because of punishment's deterrence effects.

In Figure 2, we show that wages are indeed lower in the Ex Ante treatments as compared to Baseline. We also see that of the Ex Post treatments, Strong Strategy has lower wages, while Strong Direct and Weak have wages as high as Baseline.

Table 4 confirms that these differences are significant. It appears, therefore, that principals use less wage in most cases when a powerful punishment tool—punishment with large incentive effects or that is publicly pre-committed—is available than when it is not, implying that the carrot and stick are substitutes as in Andreoni et al. (2003b).

The only exception is Strong Ex Post Direct, where wages are quite high. This result is puzzling at first, but it may be because of a greater expectation of reciprocation. When wage is interacted

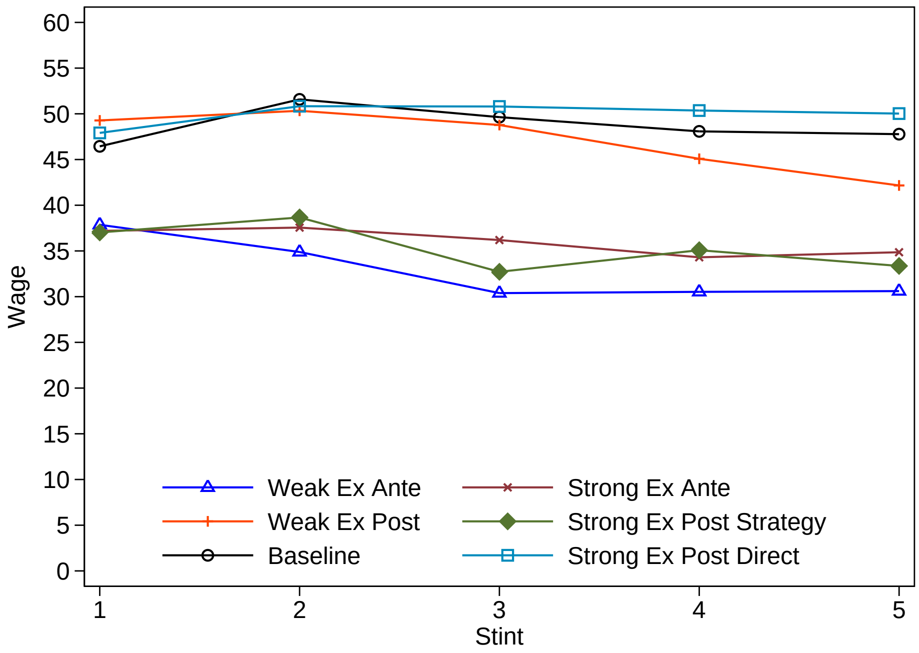
[8] Subjects were re-matched within six-person groups within their sessions. When regressions are performed at the group level instead of individual level, the only results that change are that the coefficient on punishment at effort of 0 ceases to be significant in Specification (3), round becomes significant in (3), and in (4) requested effort ceases to be significant. Also, when Tobit is used instead of OLS in individual-level regressions, all results persist. Finally, when regressions use errors clustered at the level of a six-subject group within which all pairings occurred, the only changes are that the Weak Ex Ante dummy ceases to be significant in (1).

[9] The demographic controls that are statistically significant in these regressions are: female is negative in (3); dummy for having been raised in the US is negative in (1); dummy for Black race is negative in (4); first experiment dummy is negative in (1) and (2); dummy for whether subject has taken economics classes is negative in (1) and (2); dummy for whether mother has at least a bachelor's degree is positive in (1), (2), and (3). Balance across treatments on demographic variables is largely good, but there are some small differences in age, frequency of Asian and other races, US citizens father's education, mother's education, number of economics classes, and experience in economics experiments. If demographic variables are omitted from the regressions, results do not change substantively.

Table 3: Panel OLS Regressions of Effort on Covariates

	(1)	(2)	(3)	(4)
Weak Ex Ante	-0.40**	-0.10		
	(0.15)	(0.13)		
Strong Ex Ante	0.21	0.47***	-0.18	
	(0.15)	(0.13)	(0.18)	
Weak Ex Post	-0.05	-0.05		
	(0.16)	(0.13)		
Strong Ex Post Strategy	-0.07	0.17		0.27**
	(0.16)	(0.14)		(0.14)
Strong Ex Post Direct	0.20	0.14		0.16
	(0.15)	(0.14)		(0.14)
Wage		0.02***	0.02***	0.02***
		(0.00)	(0.00)	(0.00)
Punishment for $e = 0$			0.05***	
			(0.01)	
Punishment for $e = 1$			-0.07**	
			(0.03)	
Punishment for $e = 1$ x			0.10***	
Strong Ex Ante			(0.04)	
Punishment for $e = 2$			-0.06***	
			(0.02)	
Punishment for $e = 3$			-0.01	
			(0.02)	
Requested Effort				0.11**
				(0.05)
Round	-0.02*	-0.01	-0.01	-0.02**
	(0.01)	(0.01)	(0.02)	(0.01)
Demographic Controls	Yes	Yes	Yes	Yes
Constant	1.89***	0.56	-0.25	0.39
	(0.68)	(0.64)	(1.13)	(0.81)
R^2 overall	0.08	0.25	0.38	0.26
N	1,080	1,080	360	540

Individual random effects OLS panel regressions with robust standard errors. * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$. Punishment in terms of tokens received. Demographic controls: age, gender, US citizenship, raised in the US, race dummies, experience in experiments, experience in economics classes, number of siblings, mother's and father's education, and religiosity. Omitted category is Baseline treatment for Specifications (1) and (2), Weak Ex Ante for (3), and Weak Ex Post for (4).



Vertical axis is wage, which can range from 20 to 90. Horizontal axis is stint in role.

Figure 2: Average Wage by Period Across Treatments

Table 4: Differences in Wage Across Treatments

Treatment	Wage	Different vs Baseline?	Weak & Strong Different?
Baseline	48.71 (22.87)		
Weak Ex Ante	32.86 (14.03)	0.005	
Strong Ex Ante	36.02 (17.43)	0.022	0.697
Weak Ex Post	47.13 (0.66)	0.853	
Strong Ex Post Strategy	35.37 (15.79)	0.024	0.012
Strong Ex Post Direct	49.99 (18.27)	0.550	0.322
N	36 each		

Standard deviations in parentheses. “Different vs Baseline?” and “Weak & Strong Different?” columns present *p*-values of Wilcoxon rank-sum tests.

with treatment dummies in an effort regression, none of the interaction terms are significant (results available on request; $p > 0.12$ in all cases), so wage responsiveness does not significantly differ between any punishment treatments and the Baseline. However, the point estimates of wage responsiveness are highest for the two treatments in which wage is highest (Weak Ex Ante at 0.021 and Strong Ex Post Direct at 0.028, as compared to 0.022 for Baseline), and the interaction term for Strong Ex Post Direct is greater ($p < 0.04$ in all cases) than those for the three treatments with low wages (the Ex Ante treatments and Strong Ex Post Strategy, among which it ranges between 0.013 and 0.016).

However, even this increased reciprocation does not make a higher wage a good investment for a self-interested principal. An increase in wage of 10 (which costs the principal 10) increases effort by 0.28 in Strong Ex Post Direct. An increase in effort of 1 always increases principal payoff by 30, so that an increase in effort of 0.28 increases principal payoff by 8.4. Therefore, on average, an increase in wage is not payoff-maximizing for the principal.

4.3 The Role of Punishment

Table 3 shows that punishment performs in the expected manner as an incentive device when it is used. Recall that Nash punishment profiles can elicit effort of 1 in the Weak and 2 in the Strong treatment. Specification (3) shows that across the Ex Ante treatments, punishment at efforts below Nash levels increase effort and at efforts above Nash levels never increase and sometimes decrease effort.[10] This is so even though some punishment profiles are poorly designed, as we discuss later in this section. Regression results do not change if poorly designed profiles are controlled for or omitted from the regression (results available upon request).

Table 5 shows that punishment is used in all treatments. However, it is used less when it is weaker in power, echoing studies such as Carpenter (2007), except in cases in which punishment can't incentivize higher effort (effort of 2 and 3). It also shows that it is used less when it is not pre-committed and thus is a less credible threat (except to some extent at effort of 3, where it cannot incentivize higher effort). Punishment is also not used dramatically differently between the Strategy and Direct versions of the Strong Ex Post treatment.[11]

Punishment changes the agent's best response effort at an individual level for the Ex Ante treatments. However, this punishment need not always increase effort provision for two reasons.

First, punishment can be structured to reduce best response effort. Overall 35% of profiles choose a punishment at effort of 1 that reduces best response effort, and in Strong Ex Ante 27.22% of profiles choose a punishment at effort of 2 that reduces best response effort (in the Weak treatment, punishment at effort of 2 never affects best response effort). However, the prevalence of these badly-designed punishment profiles declines across rounds (a decline from first stint to fifth stint of 34.72% to 13.89% of the time for punishment at effort of 1 and 47.22% to 13.89% for punishment at effort

[10] The net effect of punishment at effort of 1 in Strong Ex Ante is significantly positive, $p = 0.014$.

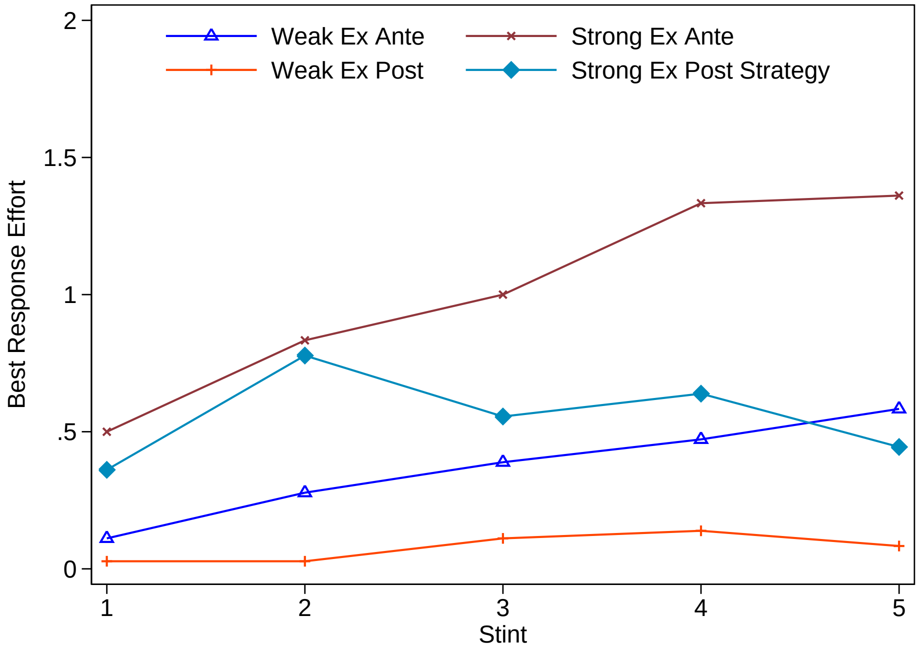
[11] Punishment at effort of 2 happens less in Direct than Strategy, and at effort of 3 it may happen less but the test is not significant because power is dramatically reduced. If the direct elicitation is easier for subjects to understand, this could imply that principals specifying punishment with the strategy method make errors, which is also implied by other results in this section.

Table 5: Average Punishment by Effort

Power	Effort = 0	Effort = 1	Effort = 2	Effort = 3
Weak Ex Ante	3.72	1.79	1.14	0.54
	(2.05)	(2.10)	(1.74)	(1.23)
Strong Ex Ante	3.82	3.17	1.12	0.31
	(2.01)	(1.96)	(1.52)	(1.06)
Weak = Strong?	0.615	0.000	0.528	0.003
Weak Ex Post	1.38	0.98	0.59	0.53
	(2.14)	(1.76)	(1.43)	(1.42)
Strong Ex Post Strategy	2.49	1.79	0.81	0.37
	(2.35)	(2.17)	(1.68)	(1.16)
Strong Ex Post Direct	2.70	1.45	0.32	0.11
	(2.27)	(2.01)	(1.01)	(0.33)
	$N = 56$	$N = 56$	$N = 59$	$N = 9$
Strategy Weak = Strong?	0.000	0.000	0.155	0.655
Direct = Strategy?	0.649	0.236	0.056	0.825
Weak: Ex Ante = Ex Post?	0.000	0.000	0.000	0.046
Strong: Ex Ante = Ex Post Strategy?	0.000	0.000	0.001	0.718

Value reported is average number of punishment tokens chosen by principal, which can range from 0 to 5. Tests reported are Wilcoxon rank-sum p-values. $N = 180$ in all cases except as noted.

of 2). Nevertheless, one of the reasons that Ex Ante punishment institutions do not perform well is because principals design punishment poorly. This could not have caused the poor performance of peer punishment in studies like Fehr et al. (1997), however, because in those papers ex ante punishment is specified as a threshold effort and an amount of punishment that will be imposed below that threshold.



Vertical axis is maximum effort that is money-maximizing best response to the chosen punishment profile; effort can range from 0 to 3. Horizontal axis is stint in role.

Figure 3: Average Best Response Effort by Period Across Treatments

More generally, however, punishment need not be used in such a way as to elicit the maximum effort theoretically achievable in a treatment. As shown in Figure 3, the effort that is best response to the principal's chosen punishment profile falls short, on average, of the Nash predictions for Ex Ante punishment. (We discuss Ex Post below.) However, in Weak and even more so in Strong Ex Ante the effort level incentivized by punishment rises over time, implying that principals learn to use punishment with experience. Principals in Weak Ex Ante also design punishment less well relative to the Nash punishment profiles than do principals in Strong Ex Ante: overall the largest effort that is best response given a punishment profile in Weak Ex Ante is 1 in 36.67% of cases and zero in the rest (i.e., zero in the majority), while in Strong Ex Ante it is 1 in 12.78% of cases and 2 in 43.89% (and zero therefore in the minority of cases). Therefore, part of the reason that punishment causes reduced effort in Weak Ex Ante is that principals do not wield it as an effective incentive, and that problem is mitigated when power is increased, particularly in later rounds. This may be in part because, since a principal must only purchase punishment tokens at the effort level the agent actually chooses, the maximum possible cost of a dracpmoam punishment profile is the same in both treatments, but the maximum benefit of a well-designed profile is twice as large in Strong Ex Ante since the effort that can be elicited is 2 instead of 1.

Second, even if punishment is well-designed, the literature (Fehr et al., 1997) has suggested that punishment that is pre-committed crowds out the desire to reciprocate and that it may even induce agents to behave spitefully (to suffer punishment to withhold gain from their principals). However, we find that effort is statistically above the best response effort in both Weak Ex Ante (Wilcoxon signed-rank test $p = 0.031$ comparing effort to largest effort that would be best response) and Strong Ex Ante ($p = 0.002$). We therefore find no evidence of a widespread spiteful response to punishment. However, agents are more likely to choose an effort below best response effort in Weak Ex Ante (21 out of 180 occasions in the population) as compared to Strong Ex Ante (14 out of 180, Wilcoxon ranksum $p = 0.011$). This difference may be another reason that principals are less driven to use effective punishment profiles in Weak Ex Ante than in Strong Ex Ante.

The case is different for Ex Post punishment. Agents can only respond to their expectation of punishment, not the punishment they will actually be subject to. Agents seem to form punishment expectations based on both the effort level requested by their current principal and their past experience with principals' punishment behavior.

Specification (4) in Table 3 shows that principals' requested effort is significantly correlated with the effort agents choose even though it is cheap talk. However, this need not be because of reciprocity. Requested effort is significantly predictive of punishment behavior. Across the Ex Post treatments, for non-zero requested effort,[12] agents are punished less at the requested effort level than at other effort levels, and this is strongly significant (two-sided un-paired t-test $p < 0.003$ in all cases) except when requested effort is 1 (when punishment at 2 and 3 is slightly but not significantly larger than that at 1, $p = 0.235$ and $p = 0.359$ respectively). Desired effort requested by principals does not differ on average across the three treatments (Wilcoxon rank-sum test $p > 0.6$ for all comparisons). Use of the requested effort signal is not all about the stick, however: wage is significantly correlated with requested effort in all Ex Post treatments, as found in Fehr and Gächter (1998). This connection is about 50% stronger in Strong Ex Post Direct (correlation coefficient 0.36 for Weak Ex Post, 0.33 for Strong Ex Post Strategy, and 0.50 for Strong Ex Post Direct).

Using an OLS panel regression (results available on request), we find that punishment specified against an agent in earlier rounds does not significantly affect behavior in Weak Ex Post (perhaps because it is too weak to strongly deter) or Strong Ex Post Direct (perhaps because information is too sparse), but does increase effort in Strong Ex Post Strategy: three points of punishment realized (i.e., one point assigned) in an earlier round is associated with an increase in effort of 0.11. However, by far the modal best response effort level in the Strategy Ex Post treatments is zero.

4.4 Efficiency

As in many existing studies (e.g. Fehr and Gächter, 2000), we find that even when peer punishment increases cooperation it is hard to find a case in which it increases social surplus. In Table 6,

[12] In the 24 cases in which requested effort is zero, there are no differences in punishment rates at the different effort levels (two-sided un-paired t-test $p > 0.376$ in all cases).

we present average gross surplus (total payoffs not counting costs of punishment) and net surplus (which reflects punishment costs) for each treatment, as well as the results of certain inter-treatment comparisons.

Table 6: Surplus Across Treatments

	Gross Surplus			**Net Surplus**		
Treatment		Different vs Baseline?	Weak & Strong Different?		Different vs Baseline?	Weak & Strong Different?
Baseline	26.32 (18.88)			26.32 (18.88)		
Weak Ex Ante	14.90 (12.94)	0.009		10.20 (13.68)	0.000	
Strong Ex Ante	32.27 (14.36)	0.116	0.000	26.78 (16.64)	0.831	0.000
Weak Ex Post	21.89 (16.25)	0.397		19.84 (16.71)	0.135	
Strong Ex Post Strategy	22.74 (11.84)	0.685	0.600	16.63 (13.76)	0.058	0.710
Strong Ex Post Direct	28.08 (15.99)	0.558	0.079	22.48 (18.88)	0.566	0.161

Numbers are in points at the principal-agent pair level. Gross Surplus is the difference between the chosen effort's benefit and cost; Net Surplus subtracts the total surplus destroyed by punishment. Standard deviations in parentheses. "Different vs Baseline?" and "Weak & Strong Different?" columns present p-values of Wilcoxon rank-sum tests. $N = 36$ for each treatment. Reported values summarize across-round averages.

Since gross surplus monotonically increases in effort, comparisons of gross surplus do not qualitatively differ from comparisons of effort levels. However, we see that Strong Ex Ante is nearly significantly more efficient than Baseline by this measure. In results not shown, we find that if subjects' first stint in a role is excluded, Strong Ex Ante does show significantly more gross surplus than Baseline (34.11 as compared to 24.47, $p = 0.022$). Considering the cost of punishment makes both Ex Ante treatments look worse: Weak shows extremely low net surplus, and the net surplus in Strong Ex Ante is very close to that of Baseline, and this result does not change if we exclude subjects' first stint. However, if we consider only agents' final stint in role, net surplus is higher in Strong Ex Ante than in Baseline (35.17 as compared to 19.44, $p = 0.020$). While this is only one stint, since it is the last stint it is suggestive that this efficiency gain might be the steady state toward which the institution drives behavior. Indeed, Gächter et al. (2008) find that while peer punishment in public good games reduces efficiency if the game is played for only 10 rounds, if the game is played for 50 rounds efficiency becomes much higher than the no-punishment baseline. In the same way, cooperation in our Strong Ex Ante punishment treatment might yield more robust and sustainable cooperation over a long period than would our no-punishment Baseline.

The net surplus values in Table 6 also reveal the cost of Ex Post punishment in our experiment.

Including the cost of punishment increases the difference between our Baseline and Weak Ex Post treatment, although we cannot reject the null hypothesis that this difference is due to chance. However, punishment significantly reduces surplus in the Strong Ex Post Strategy treatment so that it approaches the poor performance of the Weak Ex Ante treatment. This result highlights the negative qualities of both Weak Ex Ante and Strong Ex Post Strategy punishment: any increase in surplus generated by the punitive incentives is swamped by the welfare loss inherent in punishment. Strong Ex Post Direct slightly but insignificantly increases gross surplus, and slightly but insignificantly decreases net surplus. Recall that while punishment behavior was not different between Strong Ex Post Strategy and Direct, wages were higher, causing effort to be higher; this increases efficiency directly and also causes punishment to be implemented less.

The way net surplus is divided across principals and agents varies across the treatments. Principals have higher profits on average in the Strong Ex Ante (101.61 points) and Strong Ex Post Strategy (90.27) as compared to the Baseline (83.13, Wilcoxon rank-sup $p = 0.000$ and $p = 0.062$, respectively, with the test done at the subject level so $N = 36$ per cell), and lower profits in Weak Ex Post (78.35, $p = 0.072$). Agents never do significantly better than in the Baseline (63.19 points), and do significantly worse in Weak Ex Ante (47.57, $p = 0.000$), Strong Ex Ante (45.17, $p = 0.000$), and Strong Ex Post Strategy (46.37, $p = 0.000$).

4.5 Comparing Results to Public Good Literature

Why do the Strong Ex Post treatments not see gains in cooperation from peer punishment as shown in public good games in studies like Fehr and Gächter (2000)? In our results, as shown in Figure 3, punishment is not used enough to have a deterrent effect, while it is in other studies. Public good games have several players in a group rather than just two; in the case of Fehr and Gächter (2000), there are four group members, so each person faces potential punishment by three players. If we simply triple the costs of the punishment we observe in the Strong Ex Post Strategy treatment, the disincentive to choose effort of zero (at a punishment cost of 22.40 points) or one (16.10 points) would be so large that greater effort would be best response.[13] Although punishment would likely decline somewhat if more people could punish (since punishment would be a public good), the net effect of having three punishers could still be substantial.[14]

There are other differences between the public good game and the gift exchange game as we have implemented it that could alter the use and effectiveness of punishment. In particular, the asymmetry that exists in the gift exchange game could make agents perceive punishment as less "fair," and thus make them more likely to behave spitefully and less likely to be deterred from shirking; principals expecting this may punish less, and the net result may be less cooperation.

[13] The tripled punishment costs at effort of 0 are 56% of an agent's base pay from defection with a minimal wage; similarly, the punishment costs faced by the worst defectors in Fehr and Gächter (2000) are about 60% of their pay.

[14] There are two-person games in which punishment occurs after the fact and does increase cooperation, but incentives differ in them. In the prisoner's dilemma games of Bayer (2014) Tan and Xiao (2012), punishment is relatively powerful and it is a repeated game. In ultimatum and "squishy" (Rabin, 1993) games, e.g. Andreoni et al. (2003a), punishment may be quite costly, but it is quite powerful.

5 Conclusion

In many situations, people rely on each other for good behavior that can't be fully specified by contracts, and in some of these situations, peer monitoring and sanctioning is available. However, the literature has provided mixed results with regard to whether the existence of a punishment institution creates better or worse outcomes. Our study shows that some kinds of punishment can, indeed, backfire and produce lower cooperation, as found in studies like Fehr et al. (1997). This happens if punishment is weak in power and is pre-committed. But this does not seem to be primarily caused by a crowd-out of good feelings resulting from the un-trusting nature of pre-committed punishment, as some have suggested. We find little evidence of spite, and the main drivers of reduced cooperation when punishment is weak and pre-committed-to are choices made by the punisher: a lower wage (less use of a carrot when a stick is available) and punishment profiles that are not designed to incentivize high cooperation. When the same punishment is simply increased in strength, punishment increases cooperation, and this is entirely due to incentives: punishment can render a higher level of cooperation the best response, and it is better-designed by punishers, particularly as time progresses.

Indeed, in our results, the timing of punishment plays a curious role by changing the behavior of the punisher rather than the punishee. When punishment is weak in power, cooperation is lower if punishment is pre-committed than if it is chosen after the fact. This is because punishers seem to realize that weak punishment that is chosen after the fact is a poor incentive tool, so they use it little and do not lower the wage they offer when they have this kind of punishment available. When punishment is strong, cooperation is higher if punishment is pre-committed than if it is chosen after the fact, and this is because the former is used to provide incentives for high cooperation while the latter does not appear to be a credible threat—punishment that is not publicly pre-committed is simply not used enough to change behavior.

Looked at another way, our results show that some low-power punishment can reduce cooperation just as other studies like Gneezy and Rustichini (2000) have found that "small fines" can have perverse consequences, while increasing punishment's strength can correct the problem. However, in our setting, this does not seem to be caused by different framing in the mind of the punishee, but mostly by different behavior on the part of the punisher.

Finally, our results echo existing literature in finding that even when a punishment institution increases cooperation, it is difficult to find a case in which it increases net welfare (considering the costs of punishment). In fact, two of our punishment institutions show significantly worse net surplus as compared to the no-punishment baseline, and the only case in which a punishment institution significantly increases net welfare is at the very end of the interactions when punishment is high-powered and pre-committed. In other words, punishment based on incentives rather than reciprocity provides a sustainable increase in social surplus after a learning period has elapsed, but in no other case does punishment improve net welfare.

Our results help explain why punishment sometimes fails to increase cooperation and what design elements of a punishment institution render it more effective. However, we can only conjecture about why punishment based in reciprocity (i.e., decided on after the fact) is not used enough to elicit high cooperation in our study while it is in many studies in the literature (e.g. Fehr and Gächter, 2000).

6 Acknowledgements

We thank Williams College for funding. We thank seminar and conference participants and colleagues (particularly Kathy Baylis, Laura Grant, Sumeet Gulati, Corey Lang, and Lucija Muehlenbachs) for helpful comments and suggestions.

References

Akerlof, George, “Labor Contracts as Partial Gift Exchange,” *The Quarterly Journal of Economics*, 1982, *97* (4), 543–569.

Andreoni, James, Marco Castillo, and Ragan Petrie, “What do bargainers’ preferences look like? Experiments with a convex ultimatum game,” *American Economic Review*, 2003, *93* (3), 672–685.

_ , William Harbaugh, and Lise Vesterlund, “The carrot or the stick: Rewards, punishments, and cooperation,” *American Economic Review*, JUN 2003, *93* (3), 893–902.

Bayer, Ralph-C, “On the Credibility of Punishment in Repeated Social Dilemma Games,” School of Economics Working Papers 2014-08, University of Adelaide, School of Economics May 2014.

Calabuig, Vicente, Enrique Fatas, Gonzalo Olcina, and Ismael Rodriguez-Lara, “Carry a big stick, or no stick at all,” 2013.

Carpenter, Jeffrey P., “The demand for punishment,” *Journal of Economic Behavior & Organization*, APR 2007, *62* (4), 522–542.

Casari, Marco and Timothy N Cason, “The strategy method lowers measured trustworthy behavior,” *Economics Letters*, 2009, *103* (3), 157–159.

Cox, James C. and Cary A. Deck, “On the Nature of Reciprocal Motives,” *Economic Inquiry*, 2005, *43* (3), 623–635. 10.1093/ei/cbi043.

Fehr, Ernst and Armin Falk, “Psychological Foundations of Incentives,” *European Economic Review*, May 2002, *46* (4-5), 687–724.

_ and Bettina Rockenbach, “Detrimental Effects of Sanctions on Human Altruism,” *Nature*, March 2003, *422* (6928), 137–140.

_ and Simon Gächter, “How Effective are Trust-and Reciprocity-Based Incentives?,” *Economics, Values, and Organizations*, 1998, pp. 337–363.

_ and _ , “Cooperation and Punishment in Public Goods Experiments,” *American Economic Review*, September 2000, *90* (4), 980–994.

_ and _ , “Altruistic punishment in humans,” *Nature*, JAN 10 2002, *415* (6868), 137–140.

_ and _ , “Do Incentive Contracts Crowd out Voluntary Cooperation?,” Technical Report Economics Working Paper #34, Institute for Empirical Research, University of Zurich, Zurich, Switzerland 2002.

_ , Georg Kirchsteiger, and Arno Reidl, “Does fairness prevent market clearing - An experimental investigation,” *The Quarterly Journal of Economics*, MAY 1993, *108* (2), 437–459.

_ **, Simon Gächter, and Georg Kirchsteiger**, "Reciprocity as a Contract Enforcement Device: Experimental Evidence," *Econometrica*, July 1997, *65* (4), 833–860.

Fischbacher, Urs, "z-Tree: Zurich toolbox for ready-made economic experiments," *Experimental Economics*, June 2007, *10* (2), 171–178.

Gächter, Simon, Elke Renner, and Martin Sefton, "The Long-Run Benefits of Punishment," *Science*, DEC 5 2008, *322* (5907), 1510.

Gneezy, Uri and Aldo Rustichini, "A fine is a price," *The Journal of Legal Studies*, JAN 2000, *29* (1, Part 1), 1–17.

Greiner, Ben, "An Online Recruitment System for Economic Experiments," in Kurt Kremer and Volker Macho, eds., *Forschung und wissenschaftliches Rechnen 2003*, Vol. 63 of *GWDG-Bericht*, Göttingen, Germany: Gesellschaft für wissenschaftliche Datenverarbeitung mbh, 2004, pp. 79–93.

Hoffman, Elizabeth, Kevin A. McCabe, and Vernon L. Smith, "Social Distance and Other-Regarding Behavior in Dictator Games," *American Economic Review*, June 1996, *86* (3), 653–660.

Kahneman, Daniel and Shane Frederick, "Representativeness Revisited: Attribute Substitution in Intuitive Judgment," in Thomas Gilovich, Dale Griffin, and Daniel Kahneman, eds., *Heuristics and Biases: The Psychology of Intuitive Judgment*, Cambridge University Press, 2002.

Ostrom, Elinor, James Walker, and Roy Gardner, "Covenants With and Without a Sword — Self-Governance is Possible," *American Political Science Review*, JUN 1992, *86* (2), 404–417.

Putterman, Louis, *When Punishment Supports Cooperation: Insights from Voluntary Contribution Experiments*, New York, NY, USA: Oxford University Press,

Rabin, M, "Incorporating Fairness Into Game-Theory and Economics," *American Economic Review*, DEC 1993, *83* (5), 1281–1302.

Rigdon, Mary, "Trust and reciprocity in incentive contracting," *Journal of Economic Behavior & Organization*, MAY 2009, *70* (1-2), 93–105.

Tan, Fangfang and Erte Xiao, "Peer punishment with third-party approval in a social dilemma game," *Economics Letters*, 2012, *117* (3), 589 – 591.

A Session Instructions: ex ante high power

Thank you for participating in this session today! In this session, you will earn money. These instructions will explain how to earn money, so please read carefully. Before we begin, we will read the instructions together. If, after we have read through the instructions, you still have questions, please raise your hand and someone will come by to help you. Now that the experiment has begin, **we ask that you do not talk at all during the session**. Also, at this point, please turn off/silence your cell phones. Again, if you have any questions, please raise your hand and you will be addressed individually.

This session is being conducted under the Williams College Department of Economics. As per department policy, we promise there will be no deception in this session. If we were to deceive you in any way, we would be required to debrief you following the session. As there is no deception, there will be no debriefing in this session.

As you entered the room, you were given a number on a piece of paper. This number will be your ID number for the session. Your decisions will be tied only to your ID number. You will make all actions and decisions on the computer, and your decisions will be communicated to others via computer. Furthermore, your ID number will not be revealed to other subjects and other subjects will not learn what decisions and earnings you make.

This session moves step by step. No subject proceeds to the next step until all subjects complete the current step. Steps are completed by clicking a "Submit" button on the computer screen. Therefore, to keep the session moving, please do not forget to click when you are done with your decisions.

1. Introduction

This experiment lasts for ten rounds. All subjects have been randomly split into three groups. No subject knows who else is in their group. In each round, you will be randomly matched with another subject from your group, forming a 2-person pair. You can only be paired with other subjects in your group, but you will never know the identity of the person you are matched with.

Each pair consists of one Role 1 player and one Role 2 player. For each of the 5 other people in your group, you will be paired with that person for two rounds. One of these rounds you will be Role 1, and in the other round, you be will Role 2. Thus, in total, you will act as Role 1 for 5 rounds and Role 2 for 5 rounds. However, you will never be paired with the same person in the same role more than once. You will be assigned roles and pairings in random order. You will never be told who you are paired with, but you will be told which role you are at the start of each round.

Your earnings in each round will depend on the decisions of you and your match. The number of points you have at the end of each round determines your earnings from that round. After all rounds have been completed, we will randomly select one round for payment. This will be explained in further detail later.

2. Overview of the experiment

In each round, you will be making one or more decisions. At the start of each round, you will be informed whether you are Role 1 or Role 2. In each round, Role 1 will start with 100 points, while Role 2 will start with 20 points. Each round consists of 2 stages. In Stage 2, Role 2 will choose 1 of 4 actions. The actions differ both in the number of points in costs Role 2 to choose that action and in the number of points Role 1 receives if that action is chosen.

In Stage 1, Role 1 chooses how many of his tokens to transfer to his matched Role 2. Role 2 is informed of her match's Stage 1 choice when making her choice in Stage 2. These decisions will be described in further detail below.

3. Decisions

Decisions come in two stages:

Stage 1:

(Decision 1) Role 1 chooses an amount of points to transfer to Role 2. This is done by entering the amount in a box located in the center of the screen (see below).

- Transfers are taken from Role 1's points and added to Role 2's points.
- This transfer must be in whole points.
- Role 1 can choose any transfer from 20 points to 90 points.
- Role 2 will be choosing an action in the next stage; when she does so, she will see the transfer that the Role 1 player chose.
- **Example:** Role 1 transfers 35 points. Now, Role 1 has 65 points (100 minus 35), and Role 2 has 55 points (20 plus 35).

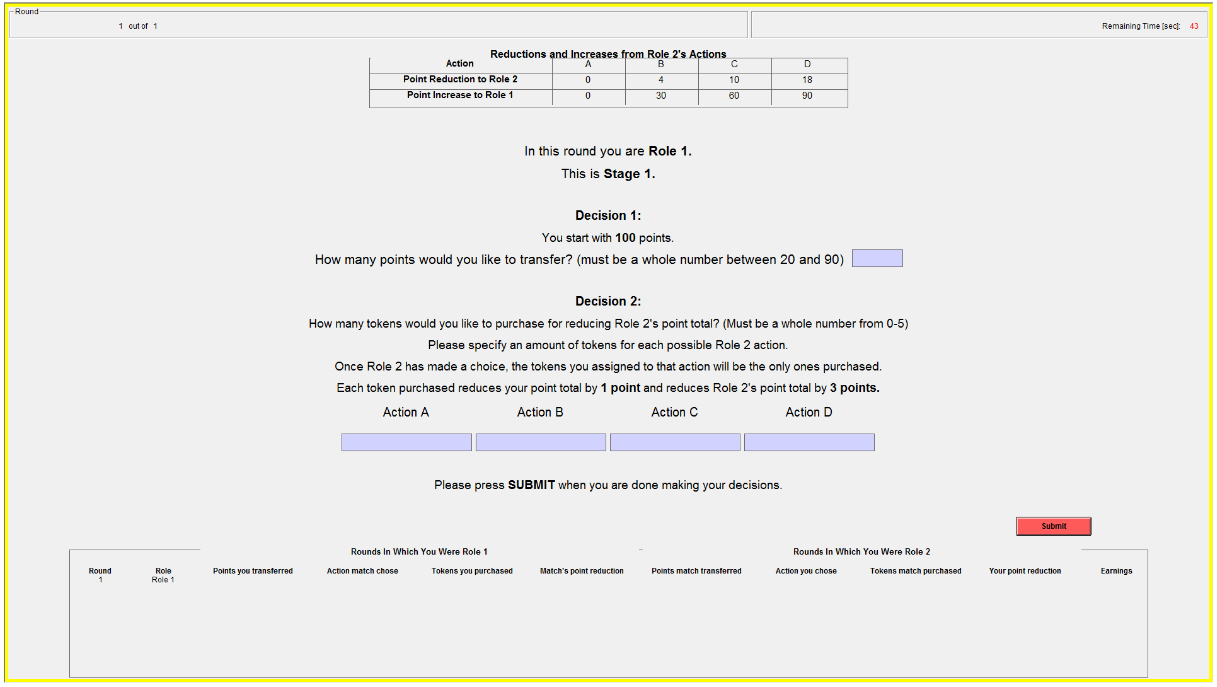
(Decision 2) In addition, Role 1 decides on a reduction profile. Role 2 will be choosing an action in Stage 2. We explain these actions below. Role 1's reduction profile indicates, for each Role 2 action, how many **tokens** Role 1 will purchase. Each token purchased reduces Role 1's point total by 1 point, and also reduces Role 2's point total by 3 points. Role 1 indicates her token decisions by filling in a table located in the center of the screen (see below).

- Role 1 must specify a how many tokens to purchase for each action Role 2 could choose (A, B, C, and D).
- Role 1 can only purchase whole tokens.
- Role 1 can purchase 0 to 5 tokens for each action.
- Role 2's point total will be reduced by 3 points for each token Role 1 purchases.
- Although Role 1 chooses a number of tokens for all 4 possible Role 2 decisions, only the tokens corresponding to Role 2's actual decision will be purchased and implemented.
- Role 2 will see the reduction profile and Role 1's transfer when she chooses her action in the next stage.

- **Example:** Example reduction profile

Action	A	B	C	D
Tokens purchased by Role 1:	2	3	0	5
Point reduction to Role 2:	6	9	0	15

Stage 1 Decision Screen:



Round

1 out of 1

Remaining Time [sec]: 43

Reductions and Increases from Role 2's Actions

Action	A	B	C	D
Point Reduction to Role 2	0	4	10	18
Point Increase to Role 1	0	30	60	90

In this round you are **Role 1.**

This is **Stage 1.**

Decision 1:

You start with **100** points.

How many points would you like to transfer? (must be a whole number between 20 and 90)

Decision 2:

How many tokens would you like to purchase for reducing Role 2's point total? (Must be a whole number from 0-5)

Please specify an amount of tokens for each possible Role 2 action.

Once Role 2 has made a choice, the tokens you assigned to that action will be the only ones purchased.

Each token purchased reduces your point total by **1 point** and reduces Role 2's point total by **3 points.**

Action A Action B Action C Action D

Please press **SUBMIT** when you are done making your decisions.

		Rounds In Which You Were Role 1				Rounds In Which You Were Role 2				
Round	Role	Points you transferred	Action match chose	Tokens you purchased	Match's point reduction	Points match transferred	Action you chose	Tokens match purchased	Your point reduction	Earnings
1	Role 1									

Notice, in both stages, the header at the top of the screen indicates what round is being played. Also, underneath the header is a table regarding Role 's actions. This table will be explained when we explain Stage 2 decisions.

At the bottom of the screen is a history table. This table displays decisions made and earnings from previous rounds. The table splits this information into the rounds in which you were Role 1, and the rounds in which you were Role 2. This table will always be at the bottom of the screen during a round.

<u>Stage 2:</u>

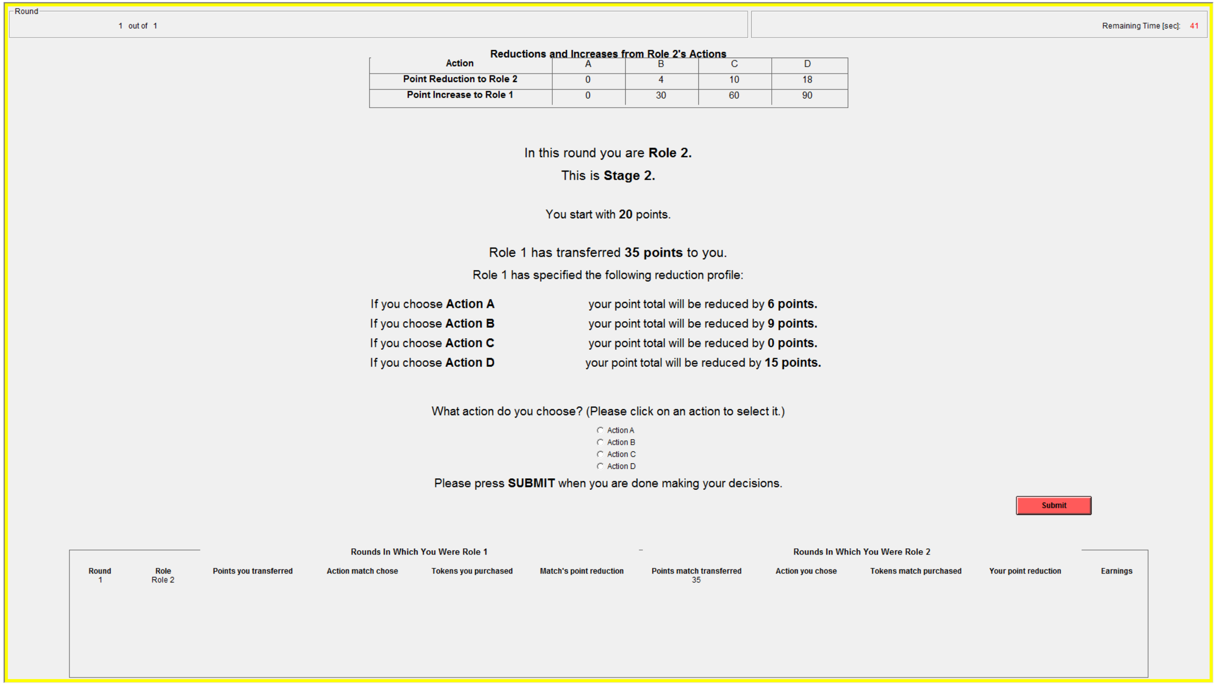
Role 2 is informed of her match's transfer and reduction profile. She must then choose an action. Her action affects both Role 1's and Role 2's point totals, as described below. Role 2 indicates her action by selecting one of four buttons provided in the lower part of the screen (see below).

- Role 2 can choose one of four actions: A, B, C, or D.
- Each action reduces Role 2's points and increases Role 1's points; the different actions correspond to different Role 2 reductions and Role 1 increases.
- The decrease to Role 2 and the increase to Role 1 for each choice is given by the following table (in points). This table will be displayed at the top of the screen whenever either player makes a decision.

Action	A	B	C	D
Point Reduction to Role 2	0	4	10	18
Point Increase to Role 1	0	30	60	90

- **Example:** Role 2 receives 35 points transferred from Role 1. Role 2 sees the reduction profile. Suppose Role 2 then chooses Action B. Role 1 gets 30 points from this choice. Role 2 loses 4 points from this choice. Role 1 purchases 3 tokens. These tokens reduce Role 2's point total by 9 points. Role 1 then has 92 points (100 minus 35 = 65, plus 30 minus 3). Role 2 then has 42 points (20 plus 35 = 55, minus 4 minus 9).

Stage 2 Decision Screen:



Reductions and Increases from Role 2's Actions

Action	A	B	C	D
Point Reduction to Role 2	0	4	10	18
Point Increase to Role 1	0	30	60	90

In this round you are **Role 2.**

This is **Stage 2.**

You start with **20** points.

Role 1 has transferred **35 points** to you.

Role 1 has specified the following reduction profile:

If you choose **Action A**	your point total will be reduced by **6 points.**
If you choose **Action B**	your point total will be reduced by **9 points.**
If you choose **Action C**	your point total will be reduced by **0 points.**
If you choose **Action D**	your point total will be reduced by **15 points.**

What action do you choose? (Please click on an action to select it.)

○ Action A
○ Action B
○ Action C
○ Action D

Please press **SUBMIT** when you are done making your decisions.

		Rounds In Which You Were Role 1				Rounds In Which You Were Role 2				
Round	Role	Points you transferred	Action match chose	Tokens you purchased	Match's point reduction	Points match transferred	Action you chose	Tokens match purchased	Your point reduction	Earnings
1	Role 2					35				

4. Earnings

At the end of each round, all decisions from the round are summarized. In addition, you will be informed of both your own and your match's earnings from that round. Earnings for each role are calculated by the following:

Role 1:

START	100 points
MINUS	(transfer to Role 2) points
PLUS	(increase from Role 2's action) points
MINUS	(reduction tokens purchased) points
EQUALS	(earnings from this round) points

Role 2:

START	20 points
PLUS	(transfer from Role 1) points
MINUS	(reduction from Role 2's action) points
MINUS	(reduction from Role 1's tokens) points
EQUALS	(earnings from this round) points

Example: For added clarity, we return to our example:

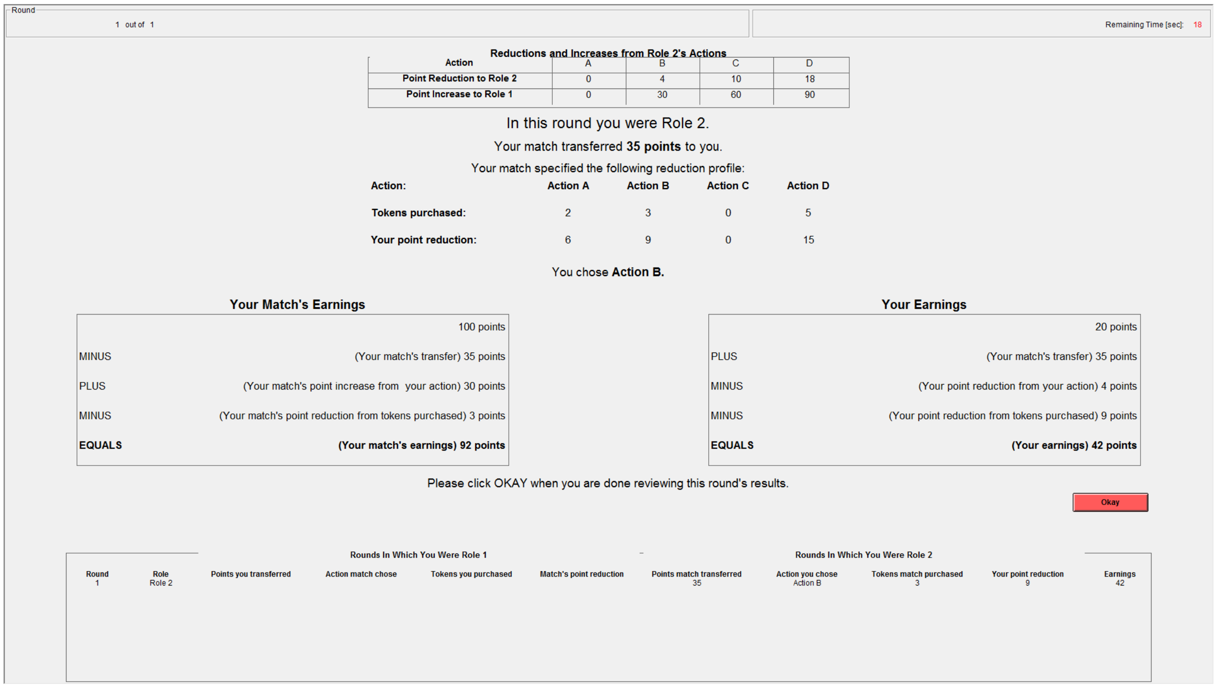
Role 1:

START	100 points
MINUS	(transfer to Role 2) 35 points
PLUS	(increase from Role 2's action) 30 points
MINUS	(reduction tokens purchased) 3 points
EQUALS	(earnings from this round) 92 points

Role 2:

START	20 points
PLUS	(transfer from Role 1) 35 points
MINUS	(reduction from Role 2's action) 4 points
MINUS	(reduction from Role 1's tokens) 9 points
EQUALS	(earnings from this round) 42 points

Review Screen:



Round

1 out of 1

Remaining Time [sec]: 18

Reductions and Increases from Role 2's Actions

Action	A	B	C	D
Point Reduction to Role 2	0	4	10	18
Point Increase to Role 1	0	30	60	90

In this round you were Role 2.

Your match transferred **35 points** to you.

Your match specified the following reduction profile:

Action:	**Action A**	**Action B**	**Action C**	**Action D**
Tokens purchased:	2	3	0	5
Your point reduction:	6	9	0	15

You chose **Action B.**

Your Match's Earnings

	100 points
MINUS	(Your match's transfer) 35 points
PLUS	(Your match's point increase from your action) 30 points
MINUS	(Your match's point reduction from tokens purchased) 3 points
EQUALS	**(Your match's earnings) 92 points**

Your Earnings

	20 points
PLUS	(Your match's transfer) 35 points
MINUS	(Your point reduction from your action) 4 points
MINUS	(Your point reduction from tokens purchased) 9 points
EQUALS	**(Your earnings) 42 points**

Please click OKAY when you are done reviewing this round's results.

		Rounds In Which You Were Role 1				Rounds In Which You Were Role 2				
Round	**Role**	**Points you transferred**	**Action match chose**	**Tokens you purchased**	**Match's point reduction**	**Points match transferred**	**Action you chose**	**Tokens match purchased**	**Your point reduction**	**Earnings**
1	Role 2					35	Action B	3	9	42

A summary of the round's decisions is displayed on a review screen. This review screen also shows how earnings were calculated for both you and your match in that round. Also, you can see previous round's decisions and earnings in the history table at the bottom of the screen. When you are done reviewing the round, do not forgot to click DONE to move on to the next round.

5. Payment

After all rounds have been completed, **one of the ten rounds will be selected at random to be the paying round for all subjects.** In front of all of you, one subject will pick a card numbered 1-10 from a deck that another subject will shuffle. Your payment will be determined only by your points in the round whose number corresponds to that card.

The instructions above explained your earnings in points. The exchange rate of points to US dollars is given by:

4 **Points** = $1 **Dollar**

Even though some actions reduce earnings of one or another subject, we have set up the experiment so that no subject can ever lose money.

Soon, you will enter your ID number into the computer. We will use your ID number to pay you. After all rounds have been completed, you will be asked to complete a brief questionnaire. While you do this, we will place each subject's earnings in an envelope marked with that subject's ID number. Then you will pick only the envelope matching your ID number. This way, your individual earnings will remain private. Once everyone has received his/her envelope, the session is completed, and you may leave.

6. Example scenarios

Prior to the ten rounds of the session, you will have to answer questions about several example scenarios. These are done to ensure that you understand how the decisions work. This quiz will be done on the computer. Every subject must answer all questions correctly before we proceed to the actual session. If you find that you have questions as you try to answer these questions, we will come to you and help you in private.