Punishment in Gift Exchange: Carrots, Sticks, Trust, and Incentives

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

March 2022

Abstract

Punishment has a mixed record in social dilemmas, and in particular the ability to punish has been found to reduce cooperation in gift exchange games. Using a lab experiment, we vary features of punishment (its strength and whether it is pre-committed) in a gift exchange game to explore how punishment is used and what its effects are. We show that punishment's effects vary with its structure in ways that depend more on the actions of the principal (punisher) than those of the agent (punishee), and that the principal uses powerful punishment to transfer surplus from the agent to herself, generally without increasing the size of the pie. We replicate the result from the literature that pre-committed peer punishment in a gift exchange game can reduce cooperation relative to a no-punishment game, but this only happens when punishment is weak. A spiteful response to threatened punishment, if it exists, is not a major driver of this reduced cooperation. Instead, principals offer lower wages when useful punishment is available, and agents lower effort reciprocally. Strong pre-committed punishment may increase cooperation, and this is because of the incentive effects of the punishment chosen and in spite of reduced wages. Punishment that is not pre-committed does not have big effects on total welfare, though if it is strong it again distributes welfare from agent to principal. How to behave in a punishment game appears to be a difficult problem that is made easier by learning: punishers often do a bad job designing pre-committed punishment but get better over time, and when after-the-fact punishment is powerful, agents learn to choose optimizing effort based on the punishment they have received. Finally, the existence of a punishment institution often decreases social surplus (net of punishment-related losses), 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

[†]The Chapin School, New York, NY; pa.aquino@gmail.com.

[‡]Department of Economics, University of Toronto, 150 St. George Street, Toronto, ON M5S 3G7; 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 many interpersonal interactions, ranging from the labor market to the hearth and home, one person (a principal) wants to get another person (an agent) to take costly action, but cannot enforce a contract and has limited incentive tools. In such cases, the agent may choose reciprocate their principal's "gift" (such as an above market wage) by exerting more effort (e.g., Fehr et al., 1993). The principal may also be able to punish the agent for low effort. A principal's ability to pay a cost to punish undesirable actions has been found to found to elicit desirable actions in certain contexts. For example, Fehr and Gächter (2000) find that the ability of peers to punish free riders can improve voluntary contributions to public goods, and Fehr et al. (1997) find that punishment can improve outcomes in principal-agent environments. However, the effectiveness of punishment in principal-agent settings is not universal, as Fehr and Gächter (2002) find that a principal's ability to punish for low effort results in both lower effort and lower surplus.

The purpose of this paper is to investigate factors contributing to the relatively poor performance of costly punishment in some gift exchange games, especially considering its success in other environments. For both the agent in a gift exchange game and a participant in a public goods game, there is a tension between her own benefit and overall surplus. While there are a myriad of differences between the two games, we consider two differences in the *punishment institutions* commonly studied which may shed light on when, if ever, punishment may be effective in the gift exchange context.

We first note that in Fehr and Gächter (2002), a participant in the public goods game could plausibly expect so much punishment for low contributions that she can increase her own payoff by increasing her contribution to the public good. In studies in which punishment backfires, the punishment available is generally weak or small (as in Fehr et al., 1997). We conjecture that this weak power is part of the reason punishment fails, since many studies have shown that punishment becomes more effective as it is stronger, as discussed in Putterman, 2014. Thus, in our study we vary the strength of the punishment available to the principal. Second, we note that whereas in the typical gift exchange game with punishment the agent knows in advance the punishment she will receive if she choose a low level of effort, a participant in a public goods game only learns her punishment (if any) after she has made her choice. It has been argued (e.g., Fehr and Falk, 2002) that incentives may crowd out intrinsic motivation, so that credible threats of punishment may backfire. In our study of the gift exchange game, we vary whether the agent knows the punishment (if any) associated with possible effort levels before making her choice. Of course, there may be an interaction between the strength of punishment and decision timing, so in our study we vary each independently.

We find that publicly-pre-committed punishment that is weak in power, as it typically has been in studies in which the backfire has been observed, does reduce cooperation relative to the no-punishment baseline. However, it's not 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 when punishment is weak and publicly pre-committed, 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. The punishment institution we study that is most effective is pre-committed and strong—so it relies on incentives and those incentives are powerful. We also show that punishers frequently fail to choose punishments designed to elicit high cooperation, and this is particularly the case when punishment is weak. Punishers seem to learn to structure these incentives properly as they play repeated rounds. Similarly, when agents are subject to after-the-fact punishment, they learn over time to choose more optimal effort. 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 very end of our treatment with strong pre-committed punishment. However, we show that when strong punishment is available, the principal on average successfully transfers surplus from the agent to herself. Overall, punishment may not be socially beneficial, but punishers use it to reduce their own costs and increase their benefits, with increasing adeptness over time.

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 are double-anonymous (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. After learning the wage, the agent chooses one of four effort levels. A higher effort levels gives a higher payment to the principal but is 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.[1] Before learning the agent's choice, the principal specifies for each effort level the number of punishment tokens she will purchase if that 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.

[1] Brandts and Charness (2011) survey the literature and note that punishment may occur less often in the strategy method, which is the opposite of what we observe, although they also see less reward with the strategy method, which we do observe.

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 actually 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 a group of six subjects within which she will play all rounds. 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.[2] 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, the subject's role and partner are randomly determined in each round.

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

[2]The fact that subjects play both roles could make them more sympathetic to their partners. 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. On the other hand, this sympathy could also increase cooperation directly. This role reversal is, however, held constant across our conditions.

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

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.

With regard to the timing of punishment we have three different conditions: Ex Ante, Ex Post Strategy, and Ex Post Direct. For the first two, we run sessions both with Weak and Strong punishment, whereas for the latter condition we run only with Strong punishment. In Ex Post Direct, the principal chooses the number of punishment tokens to purchase, if any, after learning the agent's choice. In both Ex Ante and Ex Post Strategy, the principal, without knowing the agent's choice, specifies a punishment profile: the number of punishment tokens the principal will purchase for each possible agent effort choice.

The primary distinction between Ex Ante and Ex Post is that in Ex Ante, when the agent is choosing her effort level, she knows the punishment profile chosen by principal. In the Ex Post conditions, the agent only learns principal's punishment choice *after* the agent makes her choice. While, for obvious reasons, in Ex Post Direct the agent only learns the punishment chosen by the principal for the effort actually chosen by the agent, in Ex Post Strategy, the agent learns the complete punishment profile chosen by the principal, and thus the punishment she would have received for effort levels not chosen.

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 treatments request a desired effort level from their agents. This request is cheap talk, as the experiment does not force agents to fulfill that request and principals need not (and often could not, even if this were pre-committed punishment) specify a punishment profile that would make that effort level payoff-maximizing. Principals in Ex Ante are not given a chance to send a requested effort level because the agent will see the punishment profile before making the effort choice.

3 Theory

In this section, we derive theoretical predictions for our principal-agent game configurations. We start by establishing the Subgame Perfect Nash Equilibria (SPNE) of the one-shot interaction under the assumption that subject preferences depend only on subject earnings. We then discuss how these predictions change if subjects are also driven by social preferences. 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 discuss predictions for the one-shot game.

The SPNE for self-interested agents in the Baseline (no-punishment) principal-agent game is

straightforward. The agent chooses 0 effort, and the principal pays the minimal wage (20). The equilibrium prediction is the same 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. 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 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 equilibrium even for self-interested players is that the principal will induce the agent to choose effort of at least 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). Because effort is not affected by wage, the equilibrium wage in all cases is still the minimal wage of 20.

In Ex Post Strategy, because the 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 sequential game with incomplete information, so that the only subgame is the game itself. The punishment profile the principal chooses is best response 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. The equilibria from Ex Ante are also equilibria here because they result in no punishment, and this is true of many other punishment-effort 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. We show in our results that behavior does differ, but not because of different punishment—rather, because of different wages and reciprocity. Therefore, punishment seems to be operating in the same way strategically despite the potential alternative equilibria.

In summary, with self-interested players, the predicted wage in all cases is the minimal wage. In the Baseline and the Ex Post Direct treatment, equilibrium effort is 0, with no punishment in Ex Post Direct. The most plausible equilibrium for Ex Post Strategy is also no effort and no punishment for effort of 0. In Weak Ex Ante, equilibrium effort is 1, which is supported by the principal choosing a punishment profile of ($\geq$ 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 by the principal choosing any profile that satisfies (≥ 4, ≥ 2, 0, 0).[3,4]

We now discuss how the analysis changes qualitatively if agents have preferences for anything other than monetary consequences. Most relevant are reciprocal preferences, which render agents willing to pay a monetary costs to reward those who have been "nice" and to punish those who have not (positive and negative reciprocity per the terminology of Cox and Deck, 2005). Unconditional altruism, inequality aversion, and norm enforcement could also play roles.

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 (2002) find that agents reciprocate 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 how much reciprocation is expected. Note that the principal need not have other-regarding preferences to offer a wage; she must merely expect the agent to reciprocate. If the principal does have other-regarding preferences, e.g., altruism or inequality aversion, she may offer a wage regardless of her expectations. Altruism may also drive an agent to provide effort to benefit the principal. The effect of inequality aversion on the agent's effort depends on the wage: for a small wage it would drive a lower effort, whereas for a large wage it would drive a higher effort.

Reciprocal preferences may play an additional role in the Ex Ante treatments. As in Fehr et al. (1997), agents may view pre-committed 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., reduce effort down toward the payoff-maximizing level). Agents may also 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), or perhaps inequality aversion. 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 an agent who has exerted low effort. This 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 be those who have paid generous wages. Agents who are solely self-regarding may be disciplined into giving higher effort if punishment is frequent and severe enough to increase the expected costs of low effort. Agents with other-regarding preferences may particularly wish to conform to norms

[3]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.

[4]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.

of cooperation signaled by punishment, and thus may cooperate more than is supported by the punishment they expect.

If the principal is inequality averse, either Ex Ante or Ex Post punishment would allow a principal to specify punishment that ensures that even if she provides a large wage, the outcome will not put the principal in a state of disadvantageous inequality.

In summary, altruism, inequality aversion, and trust could drive a principal to send a wage, and altruism and reciprocity (and possibly inequality aversion) could cause an agent to choose nonzero effort. With Ex Ante punishment, negative reciprocity or inequality aversion could cause the agent to work below self-interested levels. Norm enforcement, negative reciprocity, or inequality aversion could cause the principal to make even more use of Ex Ante or Ex Post punishment, and this could render even Ex Post punishment effective at increasing effort.

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.[5] Subjects were recruited through the online recruitment system ORSEE (Greiner, 2015). The experiment was programmed in z-Tree (Fischbacher, 2007). Sessions lasted 60–75 minutes. Subjects earned $17.80 on average. Subjects' average age was 19.79, and 52.78% were female.[6]

In the analysis that follows, we primarily use groups of subjects as the unit of observation. While subjects are rematched between rounds so in a sense their data are relatively independent from each other, group-level analysis is more conservative in the face of possible intra-group correlation. We also report results from individual-level analyses where those add depth. We will conduct within- and between-group univariate nonparametric tests at the group level and panel regression analysis at the group-round level.

In the following subsections, we first report how effort varies across the treatments, and then explore how wage varies and explains effort, and next how punishment is used and how it drives effort. We then study how overall efficiency and how surplus is divided varies across the treatments.

4.1 A Peer Punishment Institution Can Reduce Cooperation

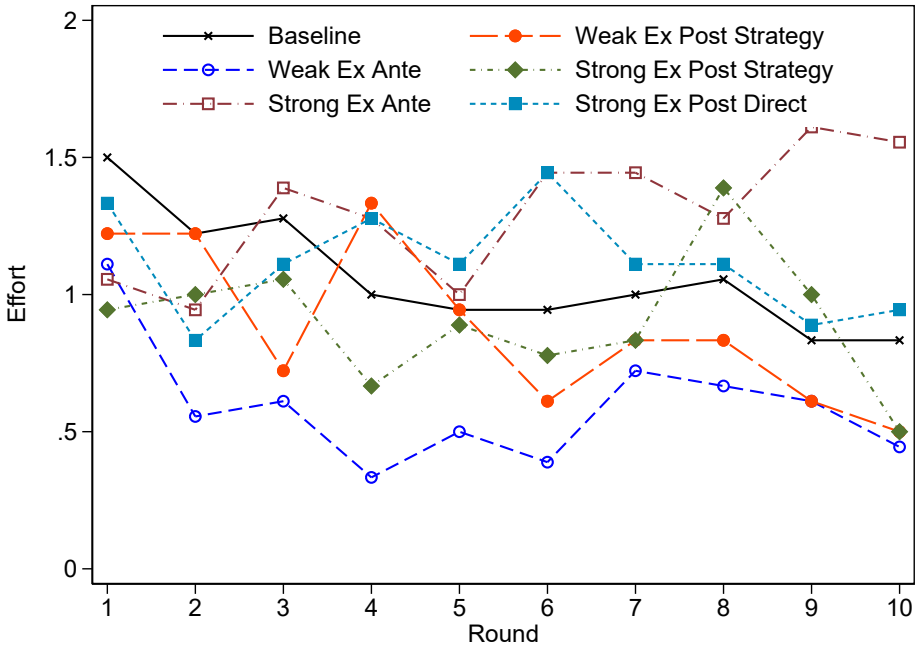
We replicate the result from Fehr et al. (1997) and others that punishment as it has typically been implemented in gift exchange games reduces cooperation. Figure 1 shows the average effort of agents across periods.[7] Weak Ex Ante shows lower effort than Baseline. Strong Ex Ante performs

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

[6] Balance across treatments on demographic variables is largely good, but there are some small differences in age, identification as Asian and other races, US citizenship, father's education, mother's education, number of economics classes, and experience in economics experiments.

[7] Different subjects are agents in different periods. Figure 1 and subsequent figures plot overall average values for each round. Trends are similar if we instead plot behavior across subjects' first, second, third, fourth, and fifth stint in the role.

better, particularly as the rounds progress, and the Ex Post treatments look similar to Baseline.



Vertical axis is effort, which can be 0, 1, 2, or 3. Horizontal axis is round.

Figure 1: Average Effort by Round Across Treatments

Table 2 shows average effort across the treatments. Weak Ex Ante's lower effort as compared to Baseline replicates the backfire of peer punishment observed in Fehr et al. (1997). This difference is on the border of marginal significance with group-level analysis; when the less-conservative individual-level analysis is conducted, it is highly significant ($p = 0.006$).

Our results show that the timing of ex ante punishment is not necessarily a recipe for worse outcomes, and we will show that when it fails it is not because pre-committed punishment causes feelings of mistrust. Effort in Ex Post is better than Ex Ante for Weak punishment (this is not significant for the group-level test but is at the individual-level, $p = 0.039$), but we show in Section 4.2 that this is explained by the higher wage offered. For Strong punishment, Ex Post actually elicits lower effort 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, not significantly for group-level tests ($p = 0.423$ compared to Weak and $p = 0.149$ to Strong) but significantly for individual-level-tests ($p = 0.056$ when compared to Weak, $p = 0.066$ when compared to Strong), but is still statistically no greater than Baseline. Since the punishment tool in the Strong Ex Post treatments is like that used in public good games like (Fehr and Gächter, 2000), it is interesting that neither provides higher effort than Baseline. We will speculate later on

Table 2: Differences in Effort Across Treatments

Treatment	Effort	Diff vs Baseline?	Weak & Strong Diff?	Ex Ante & Ex Post Diff?
Baseline	1.06			
	(0.59)			
Weak Ex Ante	0.59	0.109	0.025	0.378
	(0.44)			
Strong Ex Ante	1.30	0.521		0.077
	(0.34)			
Weak Ex Post Strategy	0.88	0.630	0.749	
	(0.52)			
Strong Ex Post Strategy	0.91	0.630		
	(0.30)			
Strong Ex Post Direct	1.12	1.000		
	(0.29)			
N	6 each			

Unit of observation is a group of six subjects. Values are average effort across subjects in the group across all rounds. Standard deviations in parentheses. "Diff" columns give p-values of Wilcoxon rank-sum tests across treatments.

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 rounds in Strong Ex Ante. If we exclude subjects' first stint in a role, individual-level analysis shows effort is significantly greater in the Strong Ex Ante treatment than Baseline (Wilcoxon rank-sum $p = 0.034$), though this is not significant in a group level analysis ($p = 0.377$). 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 that principals reduce the wage they offer when they have an effective punishment tool.

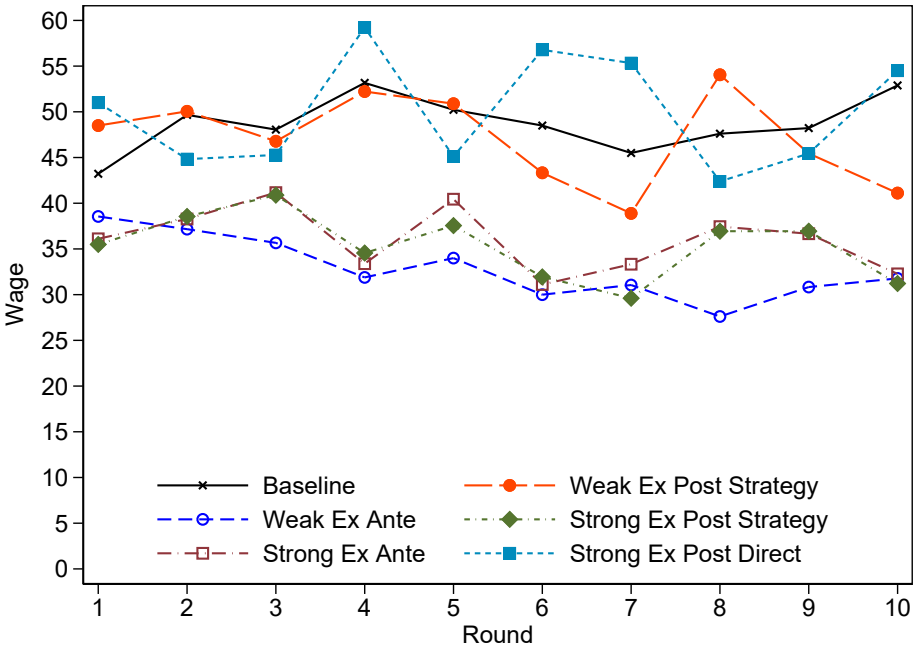
First, we demonstrate this by performing panel Tobit regressions of effort on treatment dummies and principal choices, as shown in Table 3. When regressions are performed at the individual instead of the group level, the only result that changes is that in Specification (4) requested effort is significant and positive, with a coefficient of 0.11. Also, when OLS is used instead of Tobit in group-level regressions, punishment at zero ceases to be significant.[8] We will discuss wage-related

[8]When regression is performed at the individual level and demographic controls are used, the following controls are statistically significant: female is negative in (3); the dummy for having been raised in the US is negative in (1);

results here, and we will discuss the regression results related to punishment in Section 4.3.

Specification (1) shows, as was hinted in Table 2, that Weak Ex Ante performs worse than the no-punishment Baseline. But Specification (2) shows that wage is significantly reciprocated, which has been often shown in past studies such as Fehr et al. (1997), though this may vary across populations; see Davies and Fafchamps (2021). It also shows 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 cause of reduced effort, Fehr et al. (1997) see lower wages when punishment is possible. Table 3 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.



Vertical axis is wage, which can range from 20 to 90. Horizontal axis is round.

Figure 2: Average Wage by Round Across Treatments

Table 4 confirms that many of these differences are significant or on the margin of being so. (The difference between Strong Ex Ante and Baseline is not significant in group level analysis, but it is when analysis is performed at the individual level, $p = 0.022$; similarly, Strong Ex Post Strategy is different from Baseline at the individual level of analysis, $p = 0.024$; and Weak Ex Post Strategy is different from Strong Ex Post Strategy, $p = 0.012$.) It appears, therefore, that

the first experiment dummy is negative in (1) and (2); the dummy for whether subject has taken economics classes is negative in (1) and (2); the dummy for whether mother has at least a bachelor's degree is positive in (1), (2), and (3).

Table 3: Effort as a Function of Treatment, Wage, and Punishment

	(1)	(2)	(3)	(4)
Weak Ex Ante	-0.52**	-0.22		
	(0.25)	(0.18)		
Strong Ex Ante	0.27	0.51***	-0.30	
	(0.25)	(0.17)	(0.30)	
Weak Ex Post Strategy	-0.20	-0.18		
	(0.25)	(0.17)		
Strong Ex Post Strategy	-0.12	0.13		0.33**
	(0.25)	(0.18)		(0.15)
Strong Ex Post Direct	0.09	0.06		0.23
	(0.25)	(0.17)		(0.14)
Wage		0.02***	0.02***	0.02***
		(0.00)	(0.01)	(0.00)
Punishment for $e = 0$			0.06**	
			(0.03)	
Punishment for $e = 1$			-0.12*	
			(0.06)	
Punishment for $e = 1$ x Strong Ex Ante			0.18***	
			(0.07)	
Punishment for $e = 2$			-0.10***	
			(0.03)	
Punishment for $e = 3$			0.01	
			(0.04)	
Requested Effort				-0.01
				(0.11)
Round	-0.03**	-0.02*	-0.03*	-0.03**
	(0.01)	(0.01)	(0.02)	(0.01)
Constant	1.16***	0.21	0.06	0.02
	(0.19)	(0.20)	(0.33)	(0.28)
Chi-squared	17.82	70.08	64.73	44.11
Left-censored at 0	39	39	17	14
Right-censored at 3	1	1	0	1
N	360	360	120	180

Group-level (one observation per group per round) random effects Tobit panel regressions. * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$. Punishment specified 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 Strategy for (4).

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 they treat the carrot and stick as substitutes, as in Andreoni et al. (2003).

Table 4: Differences in Wage Across Treatments

Treatment	Wage	Different vs Baseline?	Weak & Strong Different?
Baseline	48.71 (14.35)		
Weak Ex Ante	32.86 (10.10)	0.078	
Strong Ex Ante	36.02 (13.77)	0.150	0.873
Weak Ex Post Strategy	47.13 (14.75)	0.873	
Strong Ex Post Strategy	35.37 (10.07)	0.109	0.109
Strong Ex Post Direct	49.99 (8.37)	0.631	0.631
N	6 each		

Unit of observation is a group of six subjects. Values are average wage across subjects in the group across all rounds. Standard deviations in parentheses. "Different vs Baseline?" and "Weak & Strong Different?" columns give p-values of Wilcoxon rank-sum tests.

The only case in which powerful punishment and high wages coincide is Strong Ex Post Direct, where wages are quite high. Strong Ex Post Direct is significantly different from Strong Ex Post Strategy even in group-level analysis, $p = 0.025$. This is puzzling at first because of the strategic similarity between the treatments.

One hypothesis is that the high wages may arise from a greater expectation of reciprocation in the Direct elicitation treatment. However, when wage is interacted with treatment dummies in an effort regression, none of the interaction terms are significant (results available on request; $p > 0.15$ for group-level analysis and $p > 0.12$ for individual-level analysis in all cases), so wage responsiveness is not significantly greater in Strong Ex Post Direct.[9] Indeed, the ordering in the group-level analysis gives Weak Ex Post Strategy the only point estimate (0.033) of a wage responsiveness greater than Baseline (0.022), while Strong Ex Post Strategy shows the least responsiveness (0.009). In fact, in group-level analysis, wage responsiveness is significantly greater in Weak Ex Post Strategy than in all of the other treatments except Baseline, $p \leq 0.054$ in each post-estimation test.

The level of reciprocation in Weak Ex Post Strategy almost makes a higher wage a expected-payoff-neutral investment for the principal. An increase in wage of 10 (which costs the principal

[9]Individual-level analysis does show Strong Ex Post Direct to have a higher point estimate of wage responsiveness than the other treatments, but that disappears in the group-level analysis.

10) increases effort by 0.33 in Weak Ex Post Strategy. An increase in effort of 1 increases the principal's payoff by 30, so that an increase in effort of 0.33 increases principal payoff by 9.9. This means, of course, that the principal is on average losing a small amount of money, and facing a risky return, so in no case is a wage a good investment for a self-interested 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 increases effort and punishment at efforts above Nash levels never increases and sometimes decreases effort.[10] This is so even though some punishment profiles are poorly designed, as we discuss later in this section.[11]

Table 5 shows that punishment is used in all treatments and declines as effort increases. At the maximum effort, punishment is rare, and when it is used it seems likely to be an error. The results in Table 5 do not change with individual-level analysis, except that punishment at effort of 3 is significantly different between the Weak Ex Ante and Ex Post treatments, and punishment at effort of 1 is not significantly different between Strong Ex Post Direct and Strategy treatments. Punishment is used less when it is weaker, echoing studies such as Carpenter (2007), except in cases in which punishment can never incentivize higher effort (effort of 2 and 3). These results also show that punishment is used less in treatments in which 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 in any treatment). Punishment is not used dramatically differently between the Strategy and Direct versions of the Strong Ex Post treatment.[12]

Punishment can change an agent's best response effort in the Ex Ante treatments. However, this punishment need not always increase effort provision for two reasons.

First, a punishment profile can be poorly defined. In the worst cases, it can 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 (with a decline from first to fifth stint as principal 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 of 2). Poor punishment structuring could not have caused the poor performance of peer punishment in studies like Fehr et al. (1997), however,

[10] The net effect of punishment at effort of 1 in Strong Ex Ante is significantly positive, $p = 0.024$.

[11] Results of regressions run at the individual level do not change if poorly designed profiles are controlled for or omitted from the regression (results available upon request).

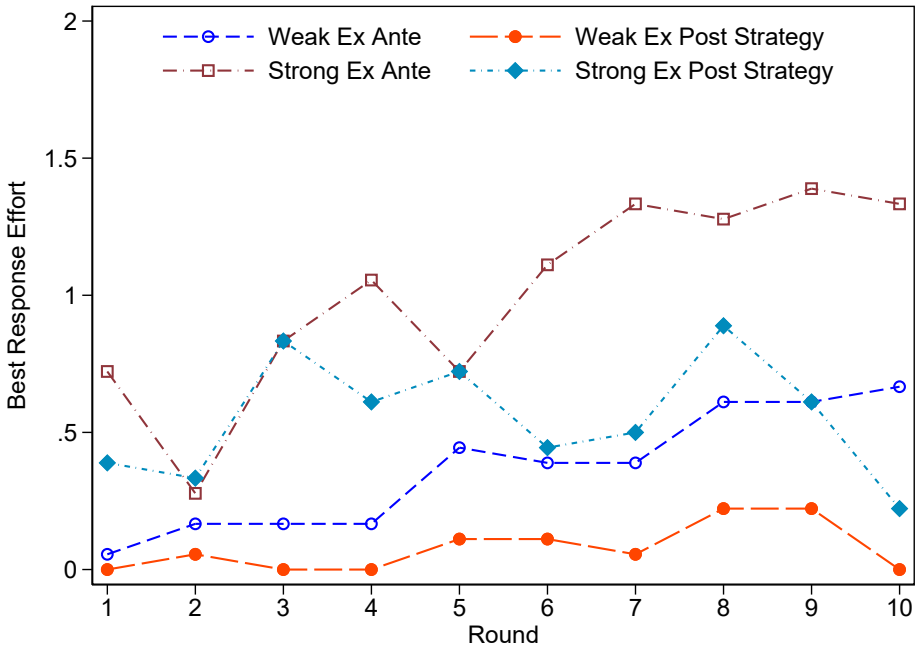
[12] Punishment at effort of 1 and 2 happens less in Direct than Strategy, and at effort of 3 it may happen less but the test is not significant because power is dramatically reduced. Power is reduced because instead of observing punishment at all possible effort levels as in the Strategy treatments, in Direct we only observe punishment in each pair at the effort level that was chosen. If the direct elicitation is easier for subjects to understand, then some of the punishment at higher effort levels in the strategy method could be the result of principals' decision-making errors; more evidence of poorly-designed punishment 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
	(1.35)	(1.29)	(1.03)	(0.70)
Strong Ex Ante	3.82	3.17	1.12	0.31
	(1.28)	(1.28)	(1.08)	(0.72)
Weak = Strong?	0.612	<0.001	0.911	0.005
Weak Ex Post Strategy	1.38	0.98	0.59	0.53
	(1.24)	(1.76)	(0.84)	(0.76)
Strong Ex Post Strategy	2.49	1.79	0.81	0.37
	(1.52)	(1.48)	(1.19)	(0.64)
Strong Ex Post Direct	2.62	1.40	0.28	0.06
	(2.08)	(1.63)	(0.68)	(0.18)
	$N = 43$	$N = 39$	$N = 37$	$N = 8$
Strategy Weak = Strong?	<0.001	0.002	0.477	0.411
Direct = Strategy?	0.749	0.054	0.004	0.209
Weak: Ex Ante = Ex Post?	<0.001	<0.001	0.001	0.509
Strong: Ex Ante = Ex Post Strategy?	<0.001	<0.001	0.013	0.336

Value reported is average number of punishment tokens chosen by principal, aggregated to the group level (i.e., one observation per group per round), which can range from 0 to 5. Tests reported are Wilcoxon rank-sum p-values. $N = 60$ in all cases except as noted.

because in those cases, 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 round.

Figure 3: Average Best Response Effort by Round Across Treatments

More generally, punishment need not be used in a way that elicits 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 well short, on average, of the Nash predictions for Ex Ante punishment. (We discuss Ex Post below.) However, especially in Strong Ex Ante, the effort level incentivized by punishment rises over time, implying that principals learn to use punishment with experience. Principals also design punishment less well relative to the Nash punishment profiles in Weak Ex Ante than 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, poor punishment design causes reduced effort in Weak Ex Ante but 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 draconian 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 may even induce agents

to behave spitefully (to suffer punishment to withhold gain from their principals). However, we find that the difference between actual effort and the maximum best response effort is 0.24 (averaged to the group level), i.e., that effort is on average above best response, and this is statistically significant (Wilcoxon signed-rank test $p = 0.017$ in Weak Ex Ante and $p < 0.001$ in Strong Ex Ante). 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), and these rates are significantly different (Wilcoxon rank-sum test at the group level $p = 0.008$). 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. However, this increased tendency to work "too little" in Weak as compared to Strong Ex Ante could be caused by either spite or error, since both should be price-sensitive.

The case is different for Ex Post punishment, where agents can only respond to their expectation of punishment, not the punishment they will actually be subject to. By far the modal best response effort level in the Strategy Ex Post treatments is zero, though of course there is variation. We will now show that agents in Ex Post treatments seem to form punishment expectations based on both the effort level requested by their current principal and their past experience with principals' punishment behavior.

First, Specification (4) in Table 3 shows that in group-level analysis, principals' requested effort is not significantly correlated with the effort agents choose; however, as noted, this correlation is positive and significant in individual-level analysis. This correlation could, but need not, be because of reciprocity. Requested effort (which does not differ significantly across treatments, Wilcoxon rank-sum test $p > 0.5$ for all comparisons) is significantly predictive of punishment behavior. Across the Ex Post treatments, for non-zero requested effort,[13] agents are punished more at effort less than as compared to effort at or above the requested effort level (group-level Wilcoxon rank-sum $p = 0.004$). Indeed, of the 173 cases in which agents work at effort levels at or above the requested effort levels, the agent is only punished in 15 (8.67%) of them. Thus, requested effort is a reasonable signal of punishment likelihood. 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 (individual-level correlation coefficient 0.36 for Weak Ex Post Strategy, 0.33 for Strong Ex Post Strategy, and 0.50 for Strong Ex Post Direct).

Next, agents can learn from past received punishment and alter their effort accordingly. Using an individual-level Tobit 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 Strategy (perhaps because it is too weak to strongly deter) but it does in Strong Ex Post Strategy (significant for both individual- and group-level analyses) and perhaps Strong Ex Post Direct (significant only for individual-level analysis, likely because information is too sparse). In

[13] In the 24 cases in which requested effort is zero, there are no differences in punishment rates at the different effort levels (individual-round-level two-sided un-paired t-test $p > 0.317$ in all cases, group-level un-paired t-tests $p > 0.292$ in all cases).

Strong Ex Post Strategy, three points of punishment realized (i.e., one point assigned) in an earlier round increase later effort by 0.9-0.11. Recall that principal-agent pairs are rematched for every interaction, so the benefits of this future good behavior are not reaped by the principal who paid for the punishment.

4.4 Efficiency

As in many existing studies (e.g., Putterman, 2014), we find that even when peer punishment increases cooperation it rarely, if ever, increases social surplus. In Table 6, 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 inter-treatment comparisons. This analysis is at the group level; results are similar when done at the individual level, except that for gross surplus the differences between Weak Ex Ante and Baseline and Strong Ex Post Direct and Weak Ex Post Strategy are significant.

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 (14.58)			26.32 (14.58)		
Weak Ex Ante	14.90 (10.62)	0.109		10.20 (10.65)	0.078	
Strong Ex Ante	32.27 (8.23)	0.522	0.016	26.78 (9.53)	1.000	0.037
Weak Ex Post Strategy	21.89 (12.62)	0.631		19.84 (12.82)	0.522	
Strong Ex Post Strategy	22.74 (7.27)	0.631	0.749	16.63 (6.84)	0.262	0.873
Strong Ex Post Direct	28.08 (7.18)	0.936	0.423	22.48 (8.01)	0.749	0.631

Numbers are in points averaged across pairs and rounds, so that the unit of observation is at the group level, with $N = 6$ per treatment. 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.

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 in individual-level analysis, 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, $p = 0.022$), but this is not significant in group-level analysis. Considering the cost of punishment makes both Ex Ante

treatments look much worse: Weak shows extremely low net surplus, and the net surplus in Strong Ex Ante is very close to that of Baseline; these results do not change if we exclude subjects' first stint. However, if we consider only agents' final stint in each role, net surplus is higher in Strong Ex Ante than in Baseline (35.17 as compared to 19.44); this is only significant at the individual-level analysis ($p = 0.020$, but $p = 0.149$ at the group level). 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 Strategy 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 a negative feature of punishment in general: any increase in surplus generated by the punitive incentives is swamped by the welfare loss that occurs if punishment is implemented. Strong Ex Post Direct slightly but insignificantly increases gross surplus, and slightly but insignificantly decreases net surplus. Recall that while punishment behavior is not different between Strong Ex Post Strategy and Direct, wages are higher in Direct, causing effort to be higher; this increases efficiency directly and also causes punishment to be implemented less. Recall that at the Nash equilibrium for rational self-interested agents, punishment is deployed in Ex Ante (but not Ex Post) treatments as an incentive but is never actually purchased, so does not reduce welfare; that is clearly not the outcome of the game as implemented here. Part of the reason is certainly the poor design of punishment profiles, as shown in the preceding section, though that cannot explain the welfare lost in the Ex Post treatments.

The way net surplus is divided across principals and agents varies across the treatments. Principals have higher profits on average in Strong Ex Ante (101.61 points) and Strong Ex Post Strategy (90.27) as compared to the Baseline (83.13, group-level Wilcoxon rank-sum $p = 0.010$ and $p = 0.078$, respectively), and lower profits in Weak Ex Post Strategy (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, $p = 0.025$), Strong Ex Ante (45.17, $p = 0.025$), and Strong Ex Post Strategy (46.37, $p = 0.025$). Therefore, although they fail to increase overall surplus, both the Strong Ex Ante and Ex Post Strategy treatments do transfer surplus from agent to principal.

5 Discussion: Comparing Results to the 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), which use a similar punishment

institution? In our results, as shown in Figure 3, ex post punishment is not used enough to have a deterrent effect, while it is in such studies. Public good games have several players in a group rather than just two; in the case of Fehr and Gächter (2000), groups have four 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 2.49 points chosen * 3 units of punishment per point * 3 punishers = 22.41 points) or one (1.79 * 3 * 3 = 16.11 points) would be so large that greater effort would be best response.[14] 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.[15]

There are other differences between gift exchange game as we have implemented it and the public good game 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, though again we don't find this to be a large driver of behavior; still, principals expecting this reaction may punish less, and the net result may be less cooperation.

6 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 they gain experience.

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

[14] 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.

[15] 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) and 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. (2003), punishment may be quite costly, but it is quite powerful.

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 rarely 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. Powerful punishment is, however, often successful at transferring surplus from agents to principals; as principals are often the ones designing institutions in the world, this may be one reason penalties are used so often even if they are not often effective.

7 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

Andreoni, James, Marco Castillo, and Ragan Petrie. 2003. "What do bargainers' preferences look like? Experiments with a convex ultimatum game." *American Economic Review*, 93(3): 672–685, DOI: http://dx.doi.org/10.1257/000282803322157034.

Andreoni, James, William Harbaugh, and Lise Vesterlund. 2003. "The carrot or the stick: Rewards, punishments, and cooperation." *American Economic Review*, 93(3): 893–902, DOI: http://dx.doi.org/10.1257/000282803322157142.

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

Brandts, Jordi, and Gary Charness. 2011. "The strategy versus the direct-response method: a first survey of experimental comparisons." *Experimental Economics*, 14(3): 375–398, DOI: http://dx.doi.org/10.1007/s10683-011-9272-x.

Carpenter, Jeffrey P. 2007. "The demand for punishment." *Journal of Economic Behavior & Organization*, 62(4): 522–542, DOI: http://dx.doi.org/10.1016/j.jebo.2005.05.004.

Cox, James C., and Cary A. Deck. 2005. "On the Nature of Reciprocal Motives." *Economic Inquiry*, 43(3): 623–635, DOI: http://dx.doi.org/10.1093/ei/cbi043.

Davies, Elwyn, and Marcel Fafchamps. 2021. "When no bad deed goes punished: Relational contracting in Ghana and the UK." *Journal of Economic Behavior and Organization*, 191 714–737, DOI: http://dx.doi.org/https://doi.org/10.1016/j.jebo.2021.09.024.

Fehr, Ernst, and Armin Falk. 2002. "Psychological Foundations of Incentives." *European Economic Review*, 46(4-5): 687–724, DOI: http://dx.doi.org/10.1016/S0014-2921(01)00208-2.

Fehr, Ernst, and Simon Gächter. 1998. "How Effective are Trust-and Reciprocity-Based Incentives?" In *Economics, Values, and Organizations*. eds. by Avner Ben-Ner, and Louis Putterman, New York, NY, USA: Cambridge University Press, 337–363.

Fehr, Ernst, and Simon Gächter. 2000. "Cooperation and Punishment in Public Goods Experiments." *American Economic Review*, 90(4): 980–994, DOI: http://dx.doi.org/10.1257/aer.90.4.980.

Fehr, Ernst, and Simon Gächter. 2002. "Do Incentive Contracts Crowd out Voluntary Cooperation?"Technical Report Economics Working Paper #34, Institute for Empirical Research, University of Zurich, Zurich, Switzerland.

Fehr, Ernst, Simon Gächter, and Georg Kirchsteiger. 1997. "Reciprocity as a Contract Enforcement Device: Experimental Evidence." *Econometrica*, 65(4): 833–860, DOI: http://dx.doi.org/10.2307/2171941.

Fehr, Ernst, Georg Kirchsteiger, and Arno Reidl. 1993. "Does fairness prevent market clearing? An experimental investigation." *The Quarterly Journal of Economics*, 108(2): 437–459, DOI: http://dx.doi.org/10.2307/2118338.

Fischbacher, Urs. 2007. "z-Tree: Zurich toolbox for ready-made economic experiments." *Experimental Economics*, 10(2): 171–178, DOI: http://dx.doi.org/10.1007/s10683-006-9159-4.

Gächter, Simon, Elke Renner, and Martin Sefton. 2008. "The Long-Run Benefits of Punishment." *Science*, 322(5907): , p. 1510, DOI: http://dx.doi.org/10.1126/Science.1164744.

Gneezy, Uri, and Aldo Rustichini. 2000. "A fine is a price." *The Journal of Legal Studies*, 29(1, Part 1): 1–17, DOI: http://dx.doi.org/10.1086/468061.

Greiner, Ben. 2015. "Subject Pool Recruitment Procedures: Organizing Experiments with ORSEE." *Journal of the Economic Science Association*, 1(1): 114–125, DOI: http://dx.doi.org/10.1007/s40881-015-0004-4.

Hoffman, Elizabeth, Kevin A. McCabe, and Vernon L. Smith. 1996. "Social Distance and Other-Regarding Behavior in Dictator Games." *American Economic Review*, 86(3): 653–660, URL: http://www.jstor.org/stable/2118218.

Putterman, Louis. 2014. "When Punishment Supports Cooperation: Insights from Voluntary Contribution Experiments." In *Reward and Punishment in Social Dilemmas*. eds. by Paul A. M. Van Lange, Bettina Rockenbach, and Toshio Yamagishi, New York, NY, USA: Oxford University Press, 17–33.

Rabin, M. 1993. "Incorporating Fairness Into Game-Theory and Economics." *American Economic Review*, 83(5): 1281–1302, URL: http://www.jstor.org/stable/2117561.

Tan, Fangfang, and Erte Xiao. 2012. "Peer punishment with third-party approval in a social dilemma game." *Economics Letters*, 117(3): 589 – 591, DOI: http://dx.doi.org/10.1016/j.econlet.2012.08.002.

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).

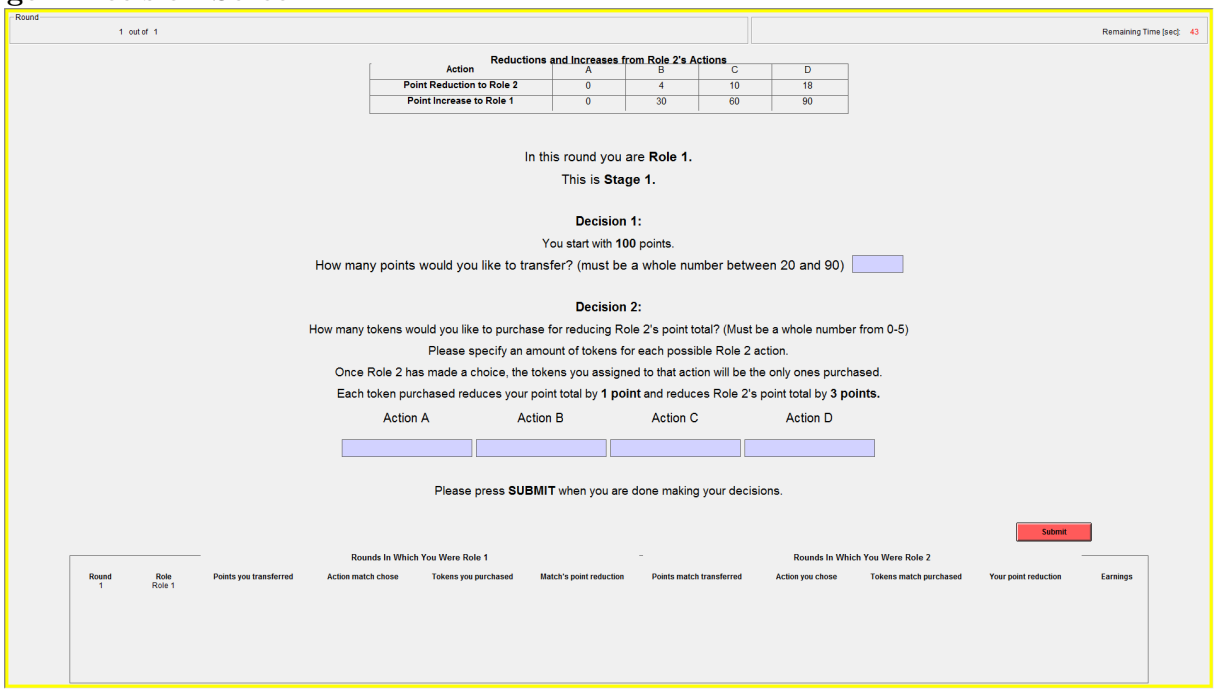
(Decision 2) In addition, Role 1 decides on a reduction profile. Role 2 will be choosing an action in Stage 2. We explain these actions below. Role 1's reduction profile indicates, for each Role 2 action, how many **tokens** Role 1 will purchase. Each token purchased reduces Role 1's point total by 1 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:



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.

Stage 2:

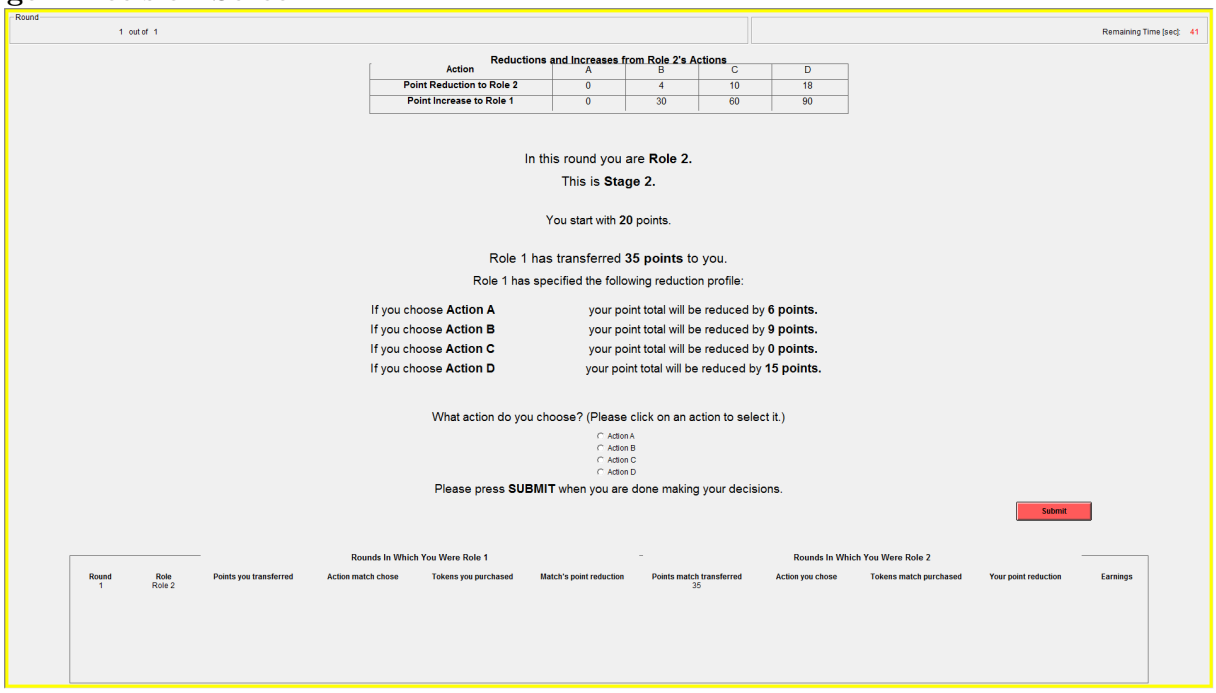
Role 2 is informed of her match's transfer and reduction profile. She must then choose an action. Her action affects both Role 1's and Role 2's point totals, as described below. Role 2 indicates her action by selecting one of four buttons provided in the lower part of the screen (see below).

- 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:



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:

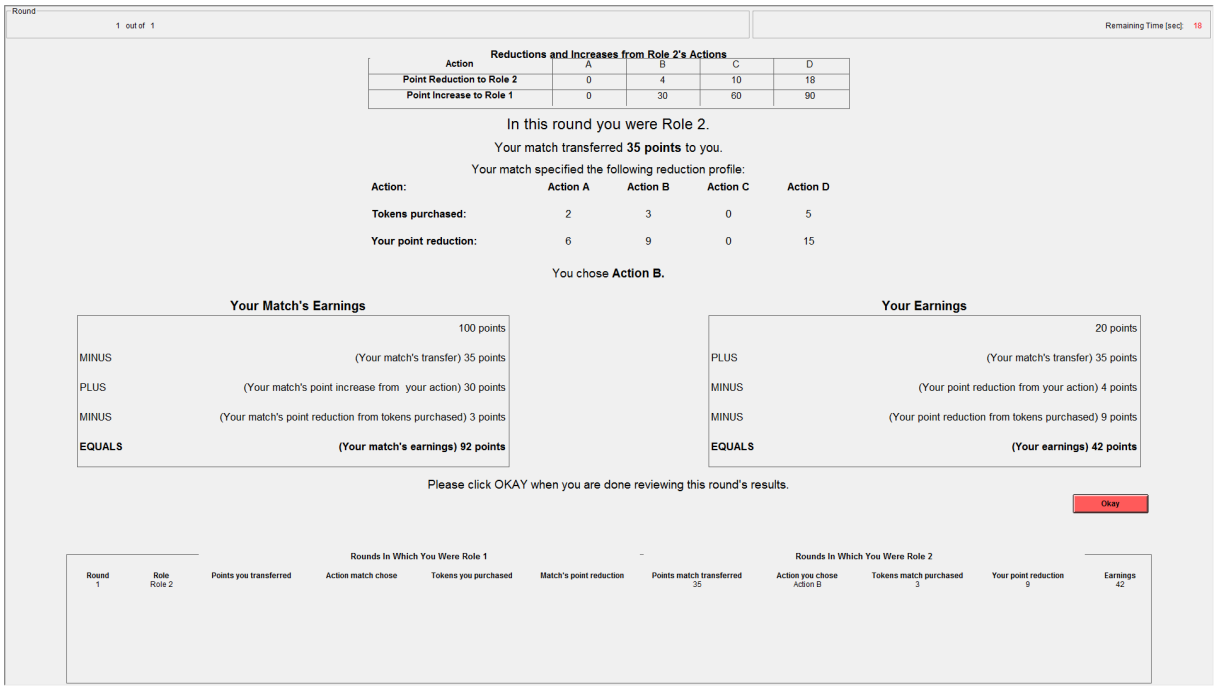
Role 1:

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:



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:

$$4 \textbf{ Points} = \$1 \textbf{ Dollar}$$

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.