Norm Enforcement with Incomplete Information

Lawrence R. De Geest*[1] and David C. Kingsley[2]

[1]Department of Economics, Suffolk University
[2]Department of Economics, University of Massachusetts Lowell

Abstract

We study the emergence of norms and their enforcement in a public goods game with private information about endowments. Subjects were randomly assigned a *Low* or *High* endowment and, across treatments, endowments were either *Observed* or *Unobserved*. In both treatments subjects could enforce contributions with peer punishment. We use punishment decisions to estimate contribution norms and then estimate the expected costs of noncompliance. In *Observed*, both *Low* and *High* types enforce a "contribute-your-endowment" norm, adjusting the costs of noncompliance to account for each type's endowment. In *Unobserved* we find that groups adapt to incomplete information by adopting a "contribute-the-*Low*-endowment" norm, and our expected cost calculations suggest that the enforcement of this norm balances the benefits of cooperation with the risk of misguided punishment. When at least one *High* type pools with *Low* types (by contributing less than or equal to the *Low* endowment), contributions of zero are punished as if they come from a *High* type, while contributions equal to the *Low* endowment are not punished in expectation (in case they come from a cooperative *Low* type). This enforcement strategy prevents cooperation from unraveling, but it also enables *High* types to hide behind the *Low* endowment. Our results dovetail with results from bargaining games and suggest that in settings with incomplete information, norms can attenuate but not eliminate non-cooperative behavior.

Keywords: Norms; Income Inequality; Incomplete Information; Cooperation; Punishment; Public Goods
JEL Codes: C92, H41, D82

*Corresponding author. Thanks for helpful comments from Brock Stoddard, Alex Smith, Darlene Chisholm, Mackenzie Alston, participants from the 2019 New England Experimental Economics Workshop and the 2019 Asia Pacific Meeting of the Economics Science Association, and three anonymous referees. Funding from University of Massachusetts Lowell is gratefully acknowledged. Declarations of interest: none.

1 Introduction

Where cooperation is valuable but cannot be contracted through economic incentives alone, norms often emerge to bridge the gap and "compensate for market failures" (Arrow, 1970). We see this in a variety of settings, from two-party exchange (e.g., spouses cooperating inside a marriage), to n-person social dilemmas (e.g., colleagues in a firm collaborating on a project or governments cooperating to mitigate pandemics and climate change).[1]

In general, a norm is a benchmark or threshold against which people can judge and sanction (or "price") each other's behavior (Vostroknutov, 2020; Fehr and Schurtenberger, 2018). This is evident in the n-person public goods game, where mutual cooperation is compromised by a freeriding incentive to each individual. In public goods experiments, groups that have a peer sanctioning mechanism and complete information about each other's behavior often establish some contribution threshold – a norm – that delineates how much they expect each other to contribute to avoid punishment (e.g., Reuben and Riedl, 2013; Fehr and Williams, 2013; Carpenter and Matthews, 2009).[2]

Yet there are many cases where cooperation is compromised not just by underlying incentives, but also by other contextual details like information asymmetries.[3] For instance, colleagues within a firm can shirk their duties by exploiting private information about their available time, just as nations can exploit private information about their ability to contribute to global public goods. Recent evidence from public goods experiments suggests that in these settings, the advantaged agent (the one with private information) will hide by matching the behavior of a cooperative, disadvantaged agent in order to avoid punishment (Kingsley, 2016; Robbett, 2016).

Can norms emerge to correct the market failures created by private information? It depends on how agents navigate the risks when peer sanctioning. Agents have to infer which group members could be hiding, so there is a risk that sanctions targeted at freeriders mistakenly hit cooperators, leading to an unraveling of cooperation (Nicklisch et al., 2016; Robbett, 2016; Ambrus and Greiner, 2012; Grechenig et al., 2010). In this paper we study how agents balance the costs of misguided punishment with the benefits of cooperation under private information, and how norms change as a result.

We study a public goods game with a fixed sanctioning institution (peer punishment) and private information. In our experiment we create private information through heterogeneous endowments. Subjects were split into groups of four. Two group members (*High* types) received a high endowment of 30 experimental dollars (EDs) and the other two a low endowment of 10 (*Low* types). In our control (*Observed*), subjects had complete information and could observe both the contribu-

[1] One example of public good provision inside firms are code review teams, in which developers are grouped to peer review updates to a codebase (Bacchelli and Bird, 2013). In terms of global public good provision, Waichman (2020) points out that peer punishment in the form of trade sanctions was used when the global community coordinated to ban chlorofluorocarbons, and that trade sanctions are regularly proposed to enforce cooperation on climate change.

[2] Norms do not have to be socially beneficial by rule. Abbink et al. (2017) provide experimental evidence of peer punishment used to promote socially-destructive norms.

[3] Reuben and Riedl (2013) note that "the contribution norm that is actually enforced may hinge on details of the environment, which may make it difficult to predict."

tions and the endowments of each group member. In our treatment (*Unobserved*), subjects had incomplete information and could only observe contributions, although they knew the distribution of endowments in their group.

In both *Observed* and *Unobserved* subjects could sanction each other with costly punishment. We use these punishment decisions to infer contribution norms in both treatments using a modified version of the Carpenter and Matthews (2009) contribution norms model. A contribution norm is simply some threshold around which groups coordinate by punishing those who contribute below it (e.g., Vostroknutov, 2020; Fehr and Schurtenberger, 2018; Reuben and Riedl, 2013; Carpenter and Matthews, 2009). Like Carpenter and Matthews (2009), we distinguish two such thresholds: one that governs the likelihood of punishment (the extensive margin of punishment) and another that governs the severity of punishment (the intensive margin of punishment). While there are many ways to study norms, estimating them from punishment decisions allows us to construct the expected cost curves for noncompliance to these norms, and thus helps us understand the enforcement strategies that emerge in different contexts. Our approach not only tells us *what* norms emerge, but importantly, *why* they emerge.[4]

The results from *Observed* replicate the main findings in the literature. Both *Low* and *High* are expected to contribute their full endowments (a norm of 10 for *Low*, 30 for *High*), just like in experiments with homogeneous endowments (Nicklisch and Wolff, 2011; Carpenter and Matthews, 2009) and with heterogeneous endowments with complete information (Reuben and Riedl, 2013). There is little ambiguity over these norms (*Low* expects *Low* to contribute 10 and *High* to contribute 30, and vice versa), and both endowment types actively participate in their enforcement. As expected, complete information helps subjects effectively establish and enforce a socially-optimal "contribute-your-endowment" norm. We show that deviations from this norm are efficiently priced for *Low* types (i.e., expected punishments overlapped with theoretical deterrence), but deviations are under-priced for *High* types.

The picture changes in *Unobserved*, where we find evidence that information shapes both norms and their enforcement. Specifically, we find that groups balance the risks of misguided punishment and the costs of freeriding by establishing a "contribute-the-*Low*-endowement" norm. We show this is accomplished through a fairly sophisticated enforcement rule that allowed punishment to vary according to what subjects knew about each other in any given round.

An important feature of our design is that it allows for endogenous information. Specifically, *High* types in any given round of *Unobserved* could either conceal (contribute 10 or less) or reveal (contribute more than 10) – but only for that round, since IDs in the punishment stage were randomized. Since we had groups of four subjects, this meant that in any given round there were

[4]In general, there are several ways to study injunctive norms (norms that dictate what action(s) people should take). Krupka and Weber (2013) measure ***beliefs*** about norms using a coordination game in which subjects independently rate the appropriateness across a set of pre-determined behaviors. Kimbrough and Vostroknutov (2016) and Kimbrough and Vostroknutov (2018) introduce rule-following tasks to measure ***propensities*** for norm compliance and show that norm sensitivity explains individual variation of pro-sociality across several experimental settings. While our focus here is to infer norms from punishment decisions, our view is that different approaches to measuring norms are complements rather than substitutes.

three possible information states: No *High* Reveal (both *High* contribute ten or less), One *High* Reveal (one *High* contributes more than ten), or Both *High* Reveal (both *High* contribute more than ten, and the group has complete information). This is a challenging environment for norms to emerge in since a subject targeting a contribution of 10 cannot discern whether that contribution came from a cooperative *Low* or a non-cooperative *High*.

We find that when at least one *High* type chooses to conceal their endowment (that is, in either the No *High* Reveal or One *High* Reveal information state), groups enforce a “contribute-the-*Low*-endowment” norm. All group members in these states are expected to contribute around 10 (the *Low* endowment), and the expected punishment at 10 is close to zero. However, the expected punishment for pure freeriding is close to the optimal punishment for *High*. In other words, subjects use punishment in a way that treats contributions of 10 as if they came from cooperative *Low*, and contributions of zero as if they came from uncooperative *High*. Moreover, *Low* types are largely responsible for this enforcement strategy, while *High* are much less involved in the maintenance of the norm (in contrast to our findings in *Observed*).

The norm and the enforcement strategy that emerges in *Unobserved* is a double-edged sword. On the one hand, it succeeds in generating positive contributions from both *Low* and *High*. On the other hand, it creates and reinforces the incentive for *High* types to hide by mimicking cooperative *Low* types, locking groups into an inferior outcome relative to *Observed*.

Our study makes several contributions. For starters, we are the first to estimate contribution norms in a public goods game with private information. In addition, our methods advance previous work on estimating norms from punishment data (Reuben and Riedl, 2013; Nicklisch and Wolff, 2011; Carpenter and Matthews, 2009) by introducing a framework for estimating the expected cost of noncompliance with contribution norms. We show that subjects adapt to private information by developing a fairly sophisticated enforcement strategy in which groups – and in particular *Low* types – promote contribution norms and enforcement that discourage pure freeriding by *High*, while at the same time avoid misguided punishment of cooperative *Low* types.

Second, we clarify the effects of incomplete information in public goods experiments with peer punishment. Bornstein and Weisel (2010) show that peer punishment is less effective when subjects have private information about their endowments, but their experiment randomly assigns endowments each period, and neither norms nor the expected cost of noncompliance are estimated. We show that when endowments are fixed over time, private information influences which norms emerge, how they are enforced, and who enforces them.

In addition, our results contrast with similar studies investigating imperfect information, such as when contributions are observed with noise (Nicklisch et al., 2016; Ambrus and Greiner, 2012; Grechenig et al., 2010). The main difference between incomplete and imperfect information is that subjects cannot exploit noise the way they can exploit private information: a *High* type cannot hide behind a small endowment if there is a chance their “hiding contribution” is flipped to a “revealing contribution”. While punishment leads to an unraveling of cooperation in games with imperfect information, we show that punishment produces stable cooperation under incomplete information, albeit less than under complete information.

Finally, our results dovetail with results from bargaining games, which suggests there is a regular pattern of behavior in strategic settings with private information and peer enforcement. In ultimatum games, splitting a surplus fifty-fifty is an established norm (Krupka and Weber, 2013; Andreoni and Bernheim, 2009), allowing first-movers with private information to "hide behind small cakes" (Güth et al., 1996; Mitzkewitz and Nagel, 1993). For example, a first-mover with a large cake can offer a fifty-fifty split of a small cake, and the second-mover accepts, unwilling to accidentally punish a fair offer. Similarly, we show that when agents in a social dilemma have private information about endowments, they too establish and enforce norms that give the benefit of the doubt to potentially cooperative behavior. At the same time, we extend this literature by showing that the opposite is true for unambiguously non-cooperative behavior: agents punish freeriding more aggressively under incomplete information than complete information.

The general pattern seems to be that agents *with* private information hide behind small cakes (or small endowments), while agents *without* private information – wary of mistakenly punishing fair offers or contributions – enforce a minimum standard of cooperative behavior. This pattern underscores the strengths and limitations of norms in strategic settings with information asymmetries. They can mitigate uncooperative behavior, but they may not be able to eliminate it, and thus may not be able to fully "complete" the incomplete contracts that abound in social life (Bowles and Hwang, 2008).

The rest of our paper proceeds as follows. Section 2 reviews the related literature on public goods with heterogeneous endowments and information asymmetries. Section 3 presents our experimental design. In Section 4 we present our results and Section 5 concludes.

2 Related Literature

In linear public goods games – where the Nash equilibrium is to contribute none of your endowment and the social optimum is to contribute all of your endowment – well-designed peer punishment (Nikiforakis and Normann, 2008) can raise contributions in the short run (Fehr and Gächter, 2000; Chaudhuri, 2011; Fehr and Schurtenberger, 2018) and payoffs in the long run (Gächter et al., 2008) when subjects have homogeneous endowments.

Carpenter and Matthews (2009) explain that punishment increases cooperation when subjects use punishment in a way that establishes a "contribute-your-endowment" norm (i.e., contribute the social optimum). By creating a benchmark, the norm clearly delineates cooperators from freeriders, and thus makes it easier for groups to allocate punishment. This is crucial because excessive or misplaced punishment (anti-social or perverse punishment targeting cooperators) can completely negate the benefits of increased contributions (Nicklisch et al., 2016; Ertan et al., 2009; Nikiforakis, 2008; Bochet et al., 2006; Cinyabugama et al., 2006).

More recently, researchers have studied peer punishment in public goods games with heterogeneous endowments. In the absence of peer punishment, endowment heterogeneity has been shown to either have no effect (Reuben and Riedl, 2013; Hofmeyr et al., 2007; Buckley and Croson, 2006) or a negative effect on cooperation (Hargreaves Heap et al., 2016; Weng and Carlsson, 2015; Cherry et al., 2005; Zelmer, 2003). The effect of introducing peer punishment into groups with heterogeneous endowments is also mixed but tends to suggest an increase in cooperation (Waichman, 2020; Visser and J., 2015; Weng and Carlsson, 2015; Reuben and Riedl, 2013).[5] One concern about endowment heterogeneity is that it could make it harder for groups to agree on contribution norms. However, Reuben and Riedl (2013) show that when subjects have complete information, linking contributions to endowments, socially-optimal contribution norms emerge.

Importantly, endowment heterogeneity creates scope for private information if subjects can observe contributions but not endowments. Few papers investigate the impact of incomplete information on the effectiveness of peer punishment when endowments are heterogeneous, and none attempt to explain how private information affects contribution norms in social dilemmas.[6] Kingsley (2016) investigates the impact of normative conflict in an interior-solution public goods design and shows that high endowment members exploit incomplete information to match the contribution of low endowment members.[7] However, Kingsley (2016) did not include a heterogeneous endowment treatment with complete information and thus could not identify the impact of incomplete information in isolation.

Another related paper is De Geest and Kingsley (2019). Using a linear public goods game, they investigate how endowment heterogeneity and incomplete information effect institutional preferences. While they similarly show that incomplete information enables high endowment members to under-contribute (rendering peer punishment less successful), they do not estimate contribution norms or the costs of noncompliance to those norms, in large part because of the short duration of their peer punishment condition (three periods).

Finally, our paper is related to Robbett (2016), who does not study endowment heterogeneity, but observes similar dynamics in a public goods game with private information. Each group member in Robbett (2016) has the same capacity to contribute towards the public good, but the payoff func-

[5]The exception here is Waichman (2020) who investigated endowment heterogeneity along with punishment effectiveness whereby it was possible that the high or low endowment members were also more effective punishers. He finds that heterogeneity in punishment effectiveness does not effect cooperation, but that endowment heterogeneity alone does reduce cooperation relative to endowment homogeneity. One notable source of variation across these experimental designs is that the punishment effectiveness (the ratio between the cost to impose and the cost of receiving punishment) varied from 1:3-1:5, where punishment increased contributions, to 1:2, where punishment did not increase contributions, consistent with the findings of Nikiforakis and Normann (2008) in homogeneous endowment public goods games.

[6]A related thread in the literature investigates the capacity of groups to self-govern under imperfect information (i.e., when subjects receive noisy signals about individual contributions). To the best of our knowledge, the literature investigating imperfect information employs homogeneous groups in linear games where individual contributions are observed with uniform noise. In such designs, the literature suggests that peer punishment has a detrimental effect on cooperation (Grechenig et al., 2010; Ambrus and Greiner, 2012; Nicklisch et al., 2016). A key difference between these games and ours is that subjects cannot strategically exploit this noisy environment in order to mimic cooperative behavior.

[7]Normative conflict implies that several, plausibly appealing, rules of behavior coexist and may therefore limit the capacity of groups to coordinate around a particular norm (Nikiforakis et al., 2012; Cappelen et al., 2007). However, the results of Reuben and Riedl (2013) suggest that normative conflict is not a significant concern in linear public goods games with endowment heterogeneity.

tion altered the incentives across *High Demanders* and *Low Demanders*. Specifically, all members have an endowment of four tokens, but their individually-rational and socially-optimal contributions differ such that the dominant strategy contribution of *High Demanders* coincides with the socially optimal contribution of *Low Demanders* on the interior of the choice set. Across treatments, group members either observed individual contributions along with member type (*Observable Incentives*) or not (*Private Incentives*). Similar to our design, private information made it difficult to differentiate cooperative and uncooperative behavior. Robbett (2016) shows that sanctions are incapable of increasing contributions in both information treatments. Of particular relevance to our study is that when incentives are private information, punishment is only targeted towards unambiguous freeriders, thus sparing from punishment *High Demanders* who match the socially optimal-contribution of *Low Demanders*.[8]

There are important variations between Robbett (2016) and the study presented here. First, in Robbett's *Private Incentives* treatment, the composition of the group remained unknown; each participant only knew that each group of three included at least one *High Demander* and one *Low Demander*. In our design, subjects knew they were in a group of two *High* endowment types and two *Low* endowment types. This creates endogenous information states in our design that we are able to exploit to uncover how groups adapt to incomplete information. Further, unlike Robbett (2016) we estimate contribution norms and the expected costs of noncompliance in order to uncover how incomplete information affects norms and their enforcement. As shown below, by estimating contribution norms across information states along with the expected costs of noncompliance we are able to show that groups adopted and enforced a "contribute-the-*Low*-endowment" norm.

3 Experimental design and methods

We study a linear public goods experiment with endowment heterogeneity and peer punishment. Our treatments vary whether subjects could observe endowments alongside individual contributions of group members. In the control, *Observed*, similar to Reuben and Riedl (2013), subjects could link individual contributions to individual endowments. In the treatment, *Unobserved*, subjects could only view contributions.

Each subject was randomly assigned a fixed endowment of experimental dollars (EDs), which they could allocate between a private account and a group account. Subjects were randomly assigned to fixed groups of four. Each group was composed of two *High* endowment members who received 30 EDs and two *Low* endowment members who received 10 EDs, and these endowments were also fixed for the entire experiment (i.e. once a *High* type always a *High* type). The distribution of endowments within a group was symmetric and known (subjects knew they were in a group of two *Low* and two *High*). The experiment lasted for 50 periods to provide ample time for contribution norms to emerge and for groups to realize the benefits of punishment (Gächter et al., 2008).

[8] In contrast, when incentives are observed, freeriding by both types is targeted, but punishment is used infrequently and its effect is weak. As noted in Robbett (2016), this result is consistent with research suggesting that sanctions are less effective in games with interior optima (Cason and Gangadharan, 2015).

The payoff to subject i with endowment k is

$$\pi_{ik} = \max \left[0, (e_{ik} - x_{ik}) + \alpha \sum_{j=1,k}^{n} x_{jk} - r \sum_{j \neq i,k}^{n-1} P_{jk,ik} \right] - c \sum_{j \neq i,k}^{n-1} P_{ik,jk} \tag{1}$$

where x_{ik} is the subject's contribution to the group account, e_{ik} is the subject's endowment, $\alpha = 0.4$ is the marginal per capita return (MPCR) from the public good, $\sum_{j=1}^{n} x_{jk}$ represents the sum of contributions to the group account from all group members, $P_{ik,jk}$ represents the number of reduction points that i imposes on other group members j at a cost of $c = 1$, and $P_{jk,ik}$ represents the number of reduction points that other group members j impose upon i at a cost of $r = 3$. With n players, $\frac{1}{n} < \alpha < 1$, and a known last period, there is a unique, symmetric Nash equilibrium where everybody freerides and contributes nothing to the public good. Similarly, there is a social optimum where subjects contribute their entire endowment to the public good.

Contribution norms. In this setting, a norm is a threshold contribution level, T, around which groups coordinate and punish those who contribute below this threshold (Vostroknutov, 2020; Fehr and Schurtenberger, 2018). Assuming a unique norm T_k emerges for each endowment, and further assuming that group members only punish deviations from these norms, the payoff function becomes

$$\pi_{ik} = \max \left[0, (e_{ik} - x_{ik}) + \alpha \sum_{j=1,k}^{n} x_{jk} - \mathbb{1}_{\{x_{ik} < T_k\}} r \sum_{j \neq i,k}^{n-1} P_{jk,ik} \right] - \mathbb{1}_{\{x_{jk} < T_k\}} c \sum_{j \neq i,k}^{n-1} P_{ik,jk} \tag{2}$$

where $\mathbb{1}$ is an indicator function. Like Carpenter and Matthews (2009) we allow T_k to be made of two components, one that governs the probability of punishment (γ_k, the extensive margin), and another for the severity of punishment (ψ_k, the intensive margin). Since payoffs are linear, the norms that generate socially optimal contributions are simply $T_k = e_k$ and thus $\gamma_k = \psi_k = e_k$ (i.e., the norm is to contribute your endowment). In our analysis we use punishment decisions to derive estimates for γ_k and ψ_k, and then use our estimates to calculate the expected the costs of noncompliance: the predicted probability of punishment times the predicted severity of punishment, for all possible deviations from the norms. (For the remainder of the paper we drop the endowment subscript k for simplicity.)

Procedure. Subjects were informed of their endowments at the start of the experiment. At the beginning of each period, each subject chose a contribution to the group account. After all contribution decisions were made, each subject was given the opportunity to punish their group members. In order to avoid excessive losses, and to ensure *High* types did not have more power in enforcement than *Low* types, subjects of all endowments were allowed to impose up to 10 reduction points per period and losses on any given period were bounded at zero unless the subject imposed punishment (Gächter et al., 2008; Reuben and Riedl, 2013).[9] The costs associated with imposing

[9]We removed power asymmetries in enforcement to focus on the effect of incomplete information in our design. However, power asymmetries are often a consequence of inequality (see for example Waichman (2020)). We explore this idea in our discussion as a topic for future research.

reduction points were referred to as *administrative costs*, while the costs associated with receiving reduction points were referred to as *reduction costs*.

In the punishment stage subjects were shown: the aggregate contribution to the group account; the individual contributions of their group members by random ID; their individual period earnings; their total earnings (equal to the sum of their individual period earnings); and a history of outcomes in previous periods. The random ID and the order of presentation of the contributions of one's group members was randomized each period to avoid reputation effects.

Endogenous information. The randomization of IDs in the punishment stage allowed *High* types to switch between revealing (contribute more than 10) and concealing (contributing up to 10). As a result, round-by-round information in *Unobserved* was endogenous. A single contribution could be linked to an endowment if that subject contributed more than 10 (thus revealing they were a *High* type). In addition, all contributions could be linked to endowments if both *High* types contributed more than 10, in which case the group had complete information (but only for that round). This created three potential information states:

1. No ***High*** reveal. Both *High* types in the group contributed 10 or less. Group members could not link any subject's contribution to their endowment.

2. One ***High*** reveal. One *High* type contributed more than 10. Group members knew that subject's endowment, but could not link the other contributions to endowments.

3. Both ***High*** reveal. Both *High* types contributed more than 10. All contributions could be linked to endowments.

It is plausible (but not necessary) that groups might enforce different norms depending on the available information. Accordingly, we control for endogenous information by estimating our norms model within each information state.

3.1 Implementation

We ran the experiment in November and December 2018 at the Cleve E. Willis experimental lab at the University of Massachusetts Amherst. We recruited subjects from the undergraduate population using ORSEE (Greiner, 2015) and implemented the experiment in z-Tree (Fischbacher, 2007). At the beginning of the experiment we passed out and read the instructions. Then we required each participant to correctly answer a set of comprehension questions before the experiment would continue.[10] Across 4 sessions, 9 and 10 groups participated in *Observed* and *Unobserved* respectively. Each session lasted about 60 minutes. On average, subjects earned $17.26, including a $7 show-up fee.

[10] Our experiment instructions are in Section D of the appendix.

4 Results

We begin by summarizing average contributions, punishments, and earnings in Section 4.1.[11] We find that contributions by *High* are significantly higher in *Observed* and, as a result, the earnings of *Low* are significantly higher in *Observed*. In *Unobserved*, we find that over half of all contributions by *High* are 10 (the *Low* endowment) or below.

In Section 4.2 we estimate the contribution norms in each treatment for *Low* and *High*. We then use our estimates to calculate the expected costs and the expected payoffs to noncompliance. In *Observed* we find a "contribute-your-endowment" norm – *Low* are expected to contribute 10, and *High* are expected to contribute 30 (similar to Reuben and Riedl, 2013; Nicklisch and Wolff, 2011; Carpenter and Matthews, 2009) – and the costs of noncompliance are close to theoretical predictions, particularly for *Low* types.

Recall that in *Unobserved*, round-by-round information is endogenous: *High* could reveal their type by contributing above 10 (or hide by contributing 10 or below), and thus change the context for other subjects when choosing punishment. It is therefore possible that different norms emerge under different information conditions. We exploit the variation across information states to understand how norms and their enforcement change as subjects adapt to private information.

Our results suggest that information does in fact influence norms and their enforcement. Further, there appears to be a clear pattern to punishment behavior. The norms and enforcement patterns that emerge suggest that subjects tried to strike a balance between preventing *High* from freeriding and rewarding *Low* for cooperating.[12]

4.1 Contributions, Punishments, and Earnings

Table 1 displays the average group contributions, punishment (sent and received), and earnings (EDs) across treatments, overall and by endowment type. The effect of information on average contributions is immediately clear, with the overall effect being driven by the behavior of *High* types. In *Observed*, contributions are significantly higher overall ($z = 1.96$, $p = 0.05$) and among *High* types ($z = 2.287$, $p = 0.022$), but statistically equivalent among *Low* types ($z = 1.43$, $p = 0.253$).

Table 1: Average Contributions, Punishment, and Earnings across treatments.

	Contributions			Earnings			Punishment Sent			Punishment Received		
	Observed	*Unobserved*	*(z)*	*Observed*	*Unobserved*	*(z)*	*Observed*	*Unobserved*	*(z)*	*Observed*	*Unobserved*	*(z)*
Pooled	14.81	9.67	1.96**	35.87	33.52	1.31	0.897	0.788	0.653	2.69	2.36	0.653
	(-5.31)	(-5.64)		(-5.97)	(-4.06)		(-0.915)	(-0.144)		(-2.75)	(-1.36)	
High	21.4	12.64	2.287**	38.37	40.41	0.653	1.14	0.598	0.00	3.32	2.35	0.327
	(-8.4)	(-8.82)		(-4.31)	(-1.51)		(-1.41)	(-0.489)		(-2.94)	(-1.25)	
Low	8.21	6.71	1.43	33.36	26.22	1.96**	0.656	0.977	1.061	2.06	2.37	1.677*
	(-2.43)	(-2.94)		(-8.48)	(-7.48)		(-0.455)	(-0.763)		(-2.7)	(-1.85)	

Means reported in Experimental Dollars. Standard deviations in parentheses.
Wilcoxon Ranksum test statistics (z) are calculated at the group level including 9 and 10 observations in *Observed* and *Unobserved*, respectively.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

[11] The code to replicate our analysis can be found at https://github.com/lrdegeest/NormEnforcement.
[12] In the appendix we explore this result with a simple evolutionary model.

Figure 1 shows average contributions over time (A and C) and the cumulative distributions of contributions (B and D). Across treatments and endowments, contributions are fairly stable after an initial learning phase in the first ten or so periods. Contributions by *High* in *Unobserved* are lower than *Observed* and far from the social optimum. We also see lower, but statistically equivalent, contributions by *Low* in *Unobserved* compared to *Observed*.

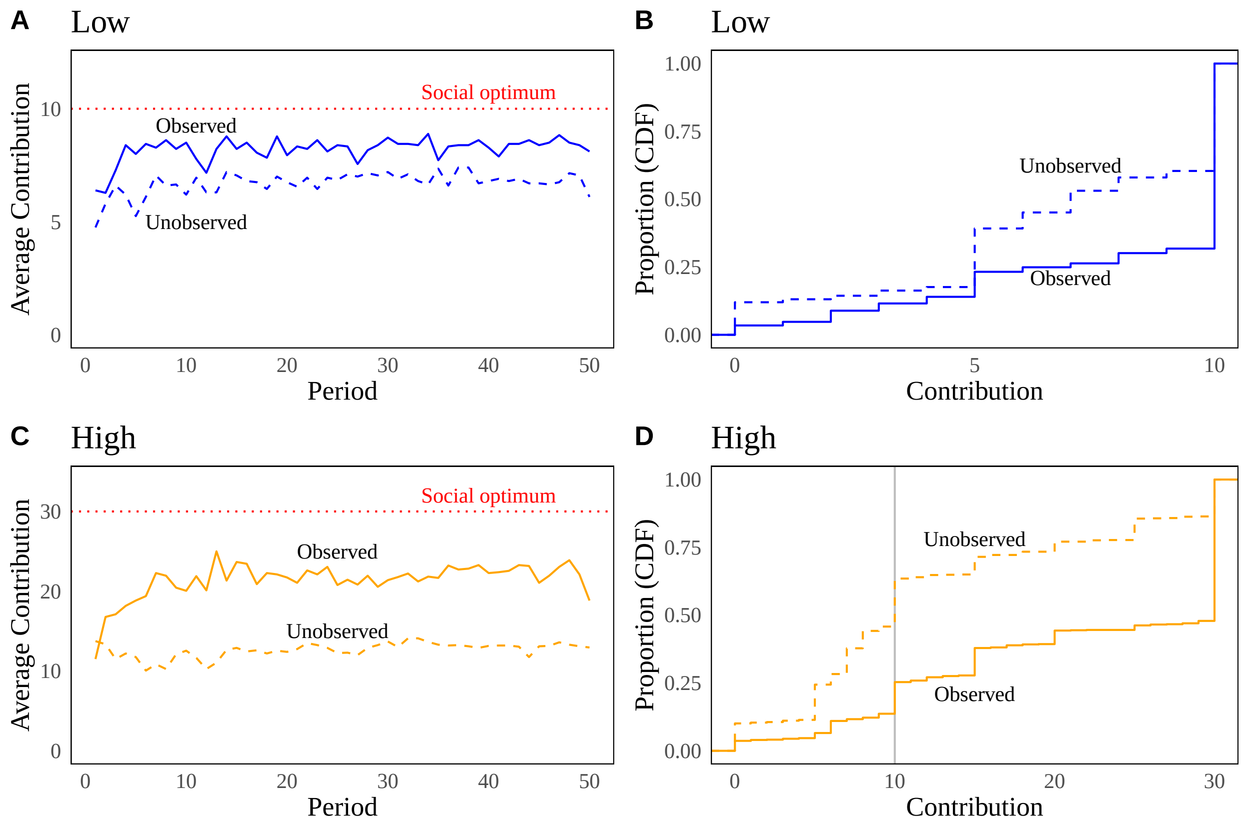


Figure 1: Panels A and C show average contributions over time. Panels B and D show the cumulative distributions of contributions. The dashed lines are *Unobserved*, the solid lines are *Observed*.

Looking at the distributions we see a higher proportion of contributions below 10 (the social optimum) by *Low* in *Unobserved*. About 68% of *Low* contributions equal the social optimum in *Observed*, compared to 40% in *Unobserved*. There is an even wider gap for *High*. About 52% of *High* contributions are exactly equal to the social optimum in *Observed*, relative to just 14% in *Unobserved*. Tellingly, about 65% of *High* contributions in *Unobserved* are 10 or less, compared to around 25% in *Observed*. Indeed, contributions by *High* in *Unobserved* are not significantly different from 10 ($z = 0.459$, $p = 0.646$).[13]

Overall, there is no significant difference in average group earnings across *Observed* and *Unobserved* ($z = 1.31$, $p = 0.19$). However, looking across endowments reveals that *Low* types in *Observed* earn significantly more than their *Unobserved* counterparts ($z = 1.96$, $p = 0.05$) while *High* types earn a statistically equivalent amount across treatments ($z = 0.653$, $p = 0.51$). Later in

[13] Wilcoxon signed-rank test with 10 group-level observations.

the paper we calculate the expected payoffs to different contributions across treatments in order to detail the variation in expected payoffs across *Unobserved* and *Observed*.

Next we turn to punishment sent and received. There is no significant difference in punishment sent across treatments overall ($z = 0.653$, $p = 0.514$), among *High* ($z = 0.00$, $p = 1.00$), or among *Low* ($z = 1.061$, $p = 0.289$). Similarly, there is no significant difference in punishment received across treatments overall ($z = 0.653$, $p = 0.514$) or among *High* ($z = 0.327$, $p = 0.744$). However, there is a small and marginally significant difference in punishment received among *Low* ($z = 1.677$, $p = 0.094$) suggesting they receive more punishment in *Unobserved* than *Observed*. However, these results combine the extensive and intensive margins of punishment and do not consider deviations from norms. In Section 4.4 we show there are substantial treatment differences in punishment when we separately estimate the margins and then calculate the expected costs of noncompliance to contribution norms.

Finally, we find that information influences participation in norm enforcement. Figure 2 shows how much each type contributes towards enforcement across treatments (Panel A) and information states in *Unobserved* (Panel B). *High* contribute most of the enforcement in *Observed*, but the roles changed in *Unobserved*, with *Low* contributing about two-thirds of all punishment, and this is stable across information states. When we estimate the treatment effect on the unconditional probability of punishment, we find a negative effect of *Observed* on *High* and a positive effect on *Low*, but the effects are insignificant after controlling for individual random effects and serial correlation within groups (*High*: $\chi^2 = 0.97, p = 0.33$; *Low*: $\chi^2 = 1.09, p = 0.30$).[14]

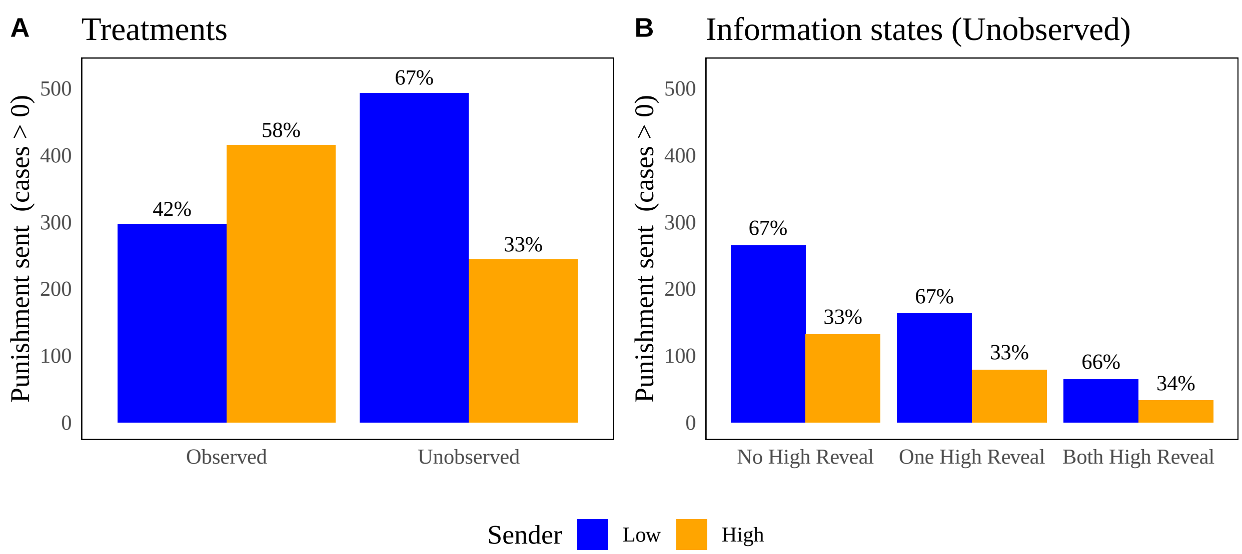


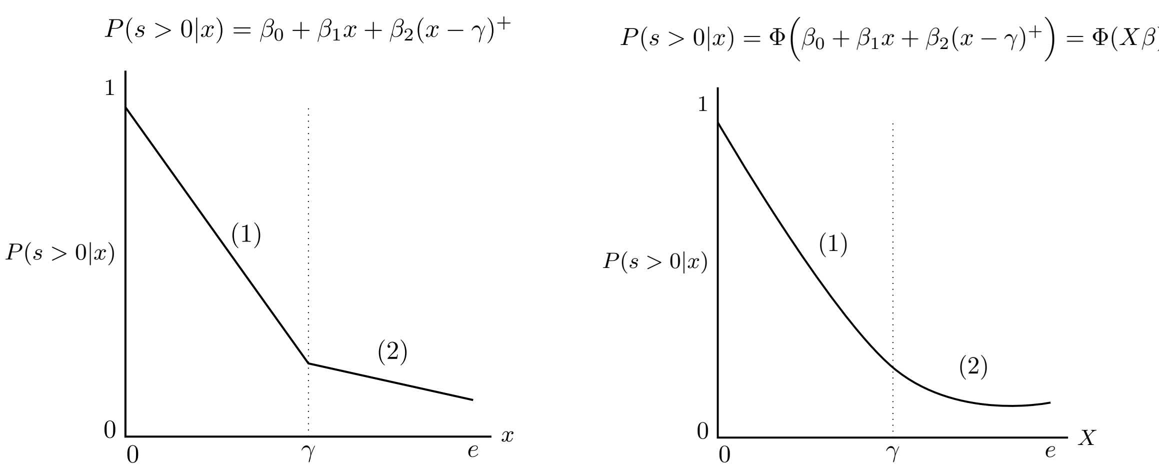
Figure 2: Contributions to enforcement by *Low* and *High* across treatments (Panel A) and information states in *Unobserved* (Panel B). The height of each bar shows the number of positive sanctions contributed by each type. The percent is that type's contribution to total enforcement. For instance, *Low* in *Observed* contributed about 300 sanctions, which was about 42% of total sanctions in that treatment.

[14]We report Wald tests on the coefficient on *Unobserved* from a random effects probit model with clustered standard errors at the group level.

4.2 Contribution norms

Contribution norms capture what subjects expected each other to contribute in order to avoid punishment. To estimate contribution norms we use a modified version of the model introduced by Carpenter and Matthews (2009).[15] Each group member i first decides whether or not to punish each other group member j (the extensive margin), and if so, how much to punish (the intensive margin).

Extensive margin. Figure 3 illustrates the basic intuition of the model on the extensive margin of punishment. Starting with a simple example, Panel A shows a linear probability model in which the probability of punishment depends only on a target's contribution x and how it compares to the contribution norm γ. In the linear model, γ is a kink or discontinuity, with β_1 the slope before the kink and $\beta_1 + \beta_2$ the slope after the kink. So, starting from γ, a marginal *decrease* in the target's contribution leads to a $|\beta_1|$ *increase* in the probability of punishment, while a marginal *increase* in the target's contribution leads to a $|\beta_1 + \beta_2|$ *decrease* in the probability of punishment, implying $|\beta_1| > |\beta_1 + \beta_2|$.



(a) Linear norms model. Marginal punishment is (1) β_1 when $x \leq \gamma$ and (2) $\beta_1 + \beta_2$ when $x > \gamma$.

(b) Nonlinear norms model. Marginal punishment is (1) $\beta_1 \phi(X\beta)$ when $x \leq \gamma$ and (2) $(\beta_1 + \beta_2)\phi(X\beta)$ when $x > \gamma$.

Figure 3: Contribution norms model for the extensive margin of punishment. Contributions below some contribution norm γ are more likely to be punished, while contributions above γ are less likely to be punished.

One problem with the linear probability model (besides generating probabilities outside the unit line) is that it predicts discrete jumps in the probability of punishment on either side of the norm. Panel 3b extends this to the nonlinear model, similar to the one used by Carpenter and Matthews (2009) and this paper, where Φ is the Normal CDF and the derivative ϕ is the Normal PDF. In this model the probability of punishment is continuous at γ, ensuring smooth predictions around the norm. Though we cannot interpret γ as a discontinuous, "hard" threshold, the intuition

[15] Reuben and Riedl (2013) also infer norms from punishments using a tobit model that collapses the extensive and intensive margins together. We used the Carpenter and Matthews (2009) model because it separately estimates the probability of punishment and the magnitude of punishment, allowing for the possibility that subject enforced different norms on each margin. This approach allows us to calculate the expected costs of noncompliance.

remains: when stable norms emerge, contributions below the norm are more likely to be punished, and contributions above the norm are less likely to be punished.

We estimate the probability of punishment using a random effects probit regression:

$$P(s > 0)_{ijgt} = \Phi\Big(\alpha + \beta_1 x_{jgt} + \beta_2 \bar{x}_{gt} + \beta_3 (x_{jgt} - \gamma)^+ + \beta_4 \bar{x}_{gt}(x_{jgt} - \gamma)^+ + \mathbf{Z}'_{igt}\psi + \mu_i + \varepsilon_{ijgt}\Big) \tag{3}$$

where $P(s > 0)_{ijgt}$ is the probability that subject i punishes subject j in group g and period t, x_{jgt} is j's contribution in t, $\bar{x}_{gt}$ is the average contribution in group g in period t, $\mathbf{Z}$ is a matrix of controls including Period, i's contribution in t and i's received sanctions in $t-1$, μ_i is the random intercept for i, and ε_{ijgt} is the idiosyncratic error. Standard errors were clustered at the group level. The term $(x_{jgt} - \gamma)^+ = \max[x_{jgt} - \gamma, 0]$ describes j's deviation above the norm γ, and thus is turned on when $x_{jgt} > \gamma$. This term allows target contributions to be treated differently on either side of the emergent norm γ. Like Carpenter and Matthews (2009) we also include the interaction term $\bar{x}_{gt}(x_{jgt} - \gamma)^+$ to control for the possibility that more cooperative groups (higher average contributions $\bar{x}_{gt}$) treat norm deviations differently than less cooperative groups.

Intensive margin. The norm on the intensive margin indicates where the severity of punishment changes. Contributions below the norm receive harsher punishments, while contributions above the norm receive milder punishments. Because sanctions were bounded below at zero and integer valued, and to account for potential non-linearity in punishment, we estimate the expected sanction size from subject i to subject j using a random effects Poisson regression:[16]

$$\mathbb{E}[s_{ijgt} | s_{ijgt} > 0] = \exp\Big(\alpha + \beta_1 x_{jgt} + \beta_2 \bar{x}_{gt} + \beta_3 (x_{jgt} - \psi)^+ + \beta_4 \bar{x}_{gt}(x_{jgt} - \psi)^+ + \mathbf{Z}'_{igt}\psi + \mu_i + \varepsilon_{ijgt}\Big) \tag{4}$$

where ψ is the contribution norm and the other covariates are the same as in Equation 3.

Procedure. To estimate the norms we run a grid search over the contribution set for each subject (e.g., $[0, 10]$ for *Low*, $[0, 30]$ for *High*); in *Unobserved* the contribution set depended on the information state (e.g., $[0, 10]$ for both *Low* and *High* in No *High* Reveal). The values of γ and ψ that maximized the likelihood of Equations 3 and 4 are interpreted as the contribution norm. Like Reuben and Riedl (2013), we collect the log-likelihood of each model and plot the normalized likelihood surface over all possible contribution norms (the worst-fitting norm is zero, and the best-fitting norm – the norm that maximizes the log-likelihood – is one). If γ (ψ) is unique then we should see a single-peaked likelihood surface, with the peak indicating the estimated norm. Otherwise the likelihood surface will be relatively flat or jagged.

[16] Carpenter and Matthews (2009) estimate the intensive margin using a linear model, which may generate predicted sanctions below zero and does not allow for non-linearity in punishment. Other studies use count data methods to estimate the intensive margin of punishment (e.g. De Geest and Stranlund, 2019).

4.3 Estimated contribution norms

Table 2 shows the estimated norms for the extensive and intensive margins in *Observed*, as well as the number of observations for each model.[17] We report the number of iterations until convergence for each likelihood-maximizing model to give a sense of the stability of our estimates.

Model (Sender $\rightarrow$ Target)	Extensive Margin (γ)	Intensive Margin (ψ)	N cases: γ (ψ)	N iterations: γ (ψ)	Log-likelihood: γ (ψ)
(1) L $\rightarrow$ L	9	9	882 (45)	6 (5)	-79.69 (-62.42)
(2) L $\rightarrow$ H	29	29	1764 (234)	6 (5)	-308.42 (-375.00)
(3) H $\rightarrow$ L	9	9	1764 (240)	5 (4)	-348.75 (-391.20)
(4) H $\rightarrow$ H	29	28	882 (158)	4 (5)	-294.72 (-246.43)

Table 2: Contribution norms in *Observed*. We estimate separate models for each sender-type and target-type (e.g., L $\rightarrow$ L means "*Low* targeting *Low*") using values for γ and ψ in the range of a target's endowment. So, *Low* targets are evaluated over the range $[0, 10]$, and *High* targets are evaluated over the range $[0, 30]$. The fourth column lists the number of observations on the extensive and intensive margins for each model. The final columns show the number of iterations it took for each model to converge and the log-likelihoods.

The results are as expected. Groups enforce a "contribute-your-endowment" norm, just like in other linear public goods games with complete information (Reuben and Riedl, 2013; Carpenter and Matthews, 2009).[18] To be precise, our model predicts that *Low* will escape punishment if they contribute 10 ($\hat{\gamma}, \hat{\psi} = 9$), and *High* if they contribute 30 ($\hat{\gamma} = 29, \hat{\psi} = 28/29$).

Interestingly, both *Low* and *High* punish the other type more frequently than they punished their own type. It is particularly stark for *Low*: they punish *High* five times more frequently than they punish *Low*. *Low* targets *Low* 5% (49/900) and *High* 14% (248/1800) of the time, while *High* targets *Low* 14% (249/1800) and *High* 18% (166/900) of the time.[19] This is more than we would expect if we adjust for the fact that *Low* could punish two *High* group members and just the one other *Low* group member.

Figure 4 shows the likelihood surface across the candidate thresholds in *Observed*. The surface for the extensive margin is single peaked, suggesting there is coordination on contribution norms (there are no other norms that come close to maximizing the log-likelihood of Equation 3). The results are noisier on the intensive margin – likely because these models are estimated using only positive values of punishment, and most punishments (about 87%) were zero.[20]

[17] In Section A of the appendix we show the estimated coefficients at the likelihood-maximizing norms.

[18] The estimated contribution norm in Carpenter and Matthews (2009) (in which subjects had identical endowments of 25) was 24. Our results are conceptually similar to Reuben and Riedl (2013). The authors use a different approach to identify contribution norms. Their free parameter, the corollary to our γ (and ψ, though their tobit model cannot separately estimate both margins of punishment), describes how much a subject expected a target to contribute as a proportion of their own contribution. Like us, they find that subjects are expected to contribute their full endowment.

[19] These numbers are slightly different than the number of observations reported in Table 2 due to the lagged variables used in the contributions norms model.

[20] This was also the case in Carpenter and Matthews (2009) and many other studies on public goods games with punishment, since punishment is a second-order public good.

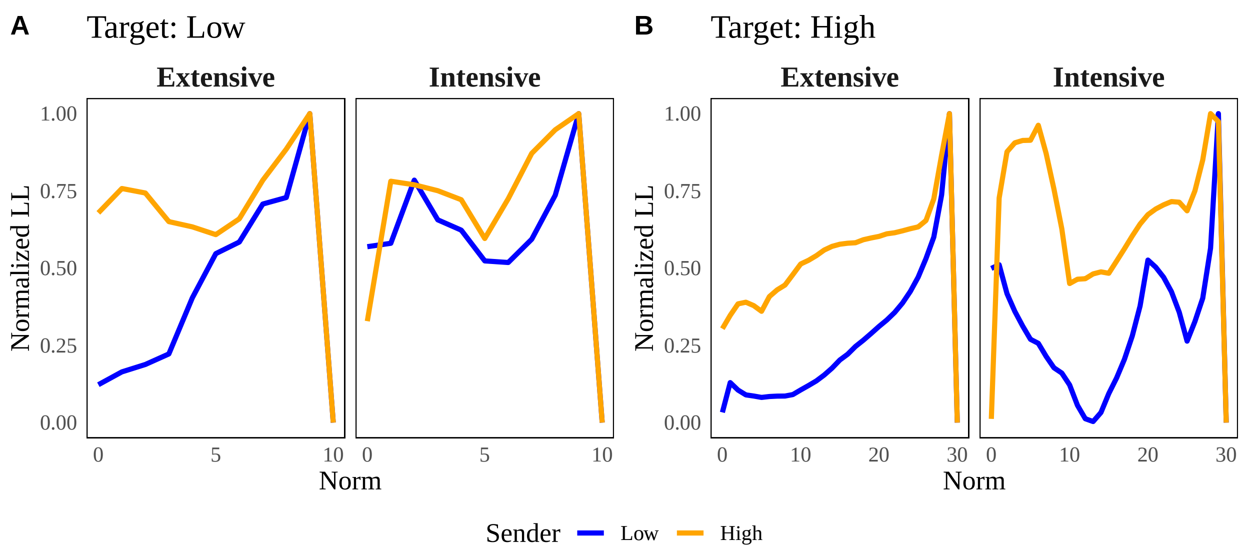


Figure 4: Likelihood surfaces for *Observed*. Likelihoods for each model (e.g., Low → Low means "*Low* targeting *Low*") are normalized so the highest value is one and the lowest is zero.

Next we turn to *Unobserved*. Table 3 shows the estimated norms across the three possible information states. Most of the models converge quickly, with a few exceptions where no unique solution is found because of too few observations. In two cases there are multiple norms on the intensive margin that maximized the likelihood.

Information State	Model (Sender → Target)	Extensive Margin (γ)	Intensive Margin (ψ)	N cases: γ (ψ)	N iterations: γ (ψ)	Log-likelihood: γ (ψ)
No *High* Reveal						
	(1) L → $[0, 10]$	8	9	1608 (262)	4 (5)	-586.93 (-424.40)
	(2) H → $[0, 10]$	1	9	1608 (131)	4 (4)	-347.74 (-227.82)
One *High* Reveal						
	(3) L → $[0, 10]$	8	8	360 (107)	3 (6)	-184.56 (-153.58)
	(4) L → H	20	[20-24]*	180 (51)	6 (7)	-109.26 (-119.57)
	(5) H → $[0, 10]$	1	7	450 (66)	6 (5)	-49.31 (-63.61*)
	(6) H → H	NA†	NA†	90 (0)	NA†	NA†
Both *High* Reveal						
	(7) L → L	1	6	264 (11)	8 (5)	-26.86 (-14.66)
	(8) L → H	20	17	528 (53)	8 (5)	-101.92 (-72.46)
	(9) H → L	2	[8-9]*	528 (24)	5 (5)	-64.83 (-40.09*)
	(10) H → H	25	NA†	264 (8)	8 (NA†)	-15.98 (NA†)

Table 3: Contribution norms in *Unobserved*. We estimate separate models for each sender-type and target-type (e.g. L → L means "*Low* targeting *Low*") and each information state (No *High* Reveal, One *High* Reveal, and Both *High* Reveal). The fifth column lists the number of observations on the extensive and intensive margins for each model. The final columns show the number of iterations it took for each model to converge and the log-likelihoods.
Notes:
*: Norms had the same log-likelihood.
†: No unique solution (model failed to converge).

The first set of models (1 and 2) estimate the contribution norms in periods where subjects could not link any contributions to endowments, No *High* Reveal. All contributions in this state are less than or equal to 10. This is the most common information state in our data and represents 55% of all punishment activity.

Figure 5 shows the likelihood surfaces for models 1 and 2. The surfaces are mostly single-peaked, particularly for *Low* senders, suggesting that the likelihood-maximizing norms are unique and fairly stable. In this information state, the model predicts that *Low* senders enforce a norm of around 10 (the *Low* endowment) on both margins. *High* senders enforce the same norm on the intensive margin, but they enforce a lower threshold of 2 on the extensive margin. As noted earlier, we find that *Low* types more actively enforce norms in this state. *Low* target other group members 16% (265/1614) of the time relative to *High* members, who target others only 8% (132/1614) of the time.

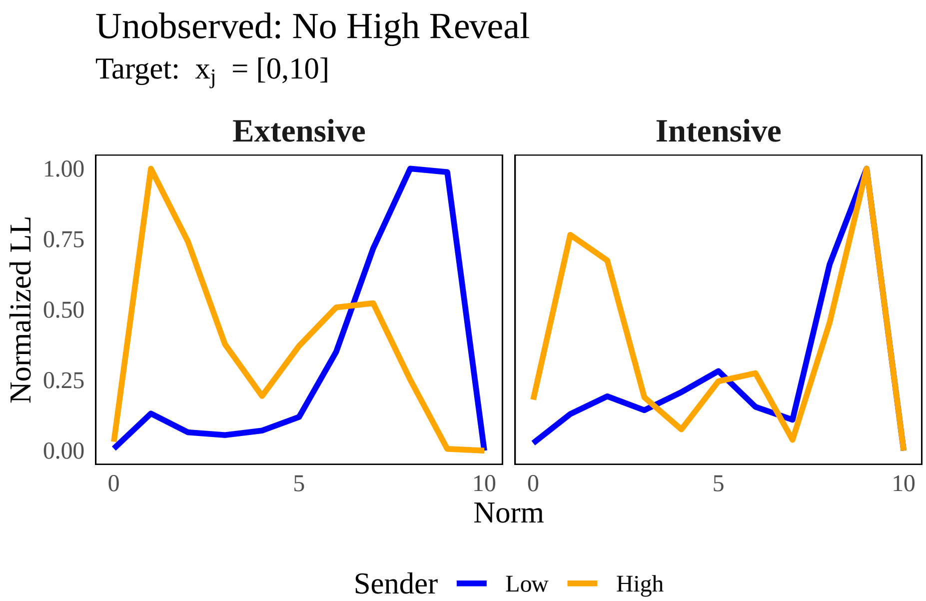


Figure 5: Likelihood surfaces for *Unobserved* in No *High* Reveal. Subjects in this information state could not link the contribution of target j to their endowment because all contributions were between 0 and 10.

The next set of models (3 to 6) estimate norms in One *High* Reveal. Here, subjects can identify one group member as *High*, but the other *High* is concealed among the two *Low*. We estimate separate models (3 and 5) on contributions between 0 and 10, and models (4 and 6) on contributions that revealed *High*. For contributions of 10 or less in this state, we find similarities with No *High* Reveal. Specifically, *Low* enforce a norm close to the *Low* endowment on both margins while *High* continued to enforce a similar norm on the intensive margin but a lower norm on the extensive margin. Again, *Low* are more active punishers in this state, targeting 28% (110/388) of the time relative 16% (76/485) of the time for *High*.

For contributions above 10, the norms enforced by *Low* and *High* types further diverge from each other and from those estimated in *Observed*. While *Low* frequently target the revealed *High*, 29% (53/94) of the time, they appear to lower their expectations of cooperation. Where *Low* enforced a "contribute-your-endowment" norm in *Observed* ($\hat{\gamma}, \hat{\psi} = 29$ on both margins), here they enforce a norm of 21 on the extensive margin ($\hat{\gamma} = 20$) and between 21-25 on the intensive margin ($\hat{\psi} = [20, 24]$). In contrast, *High* so infrequently target contributions above 10 – only 3% (3/97) of the time – that neither the extensive nor intensive margin could be estimated.

We see the starkest differences in contribution norms in *Unobserved* compared to *Observed* in the final information state, Both *High* Reveal. Recall that in this information state, subjects in *Unobserved* can link all contributions to endowments and thus had complete information (just as in *Observed*).

Starting with *Low* targeting *Low* (model 7), we find they enforce lower contribution thresholds on the extensive margin ($\hat{\gamma} = 1$) and on the intensive margin ($\hat{\gamma} = 6$), alongside a sharp drop in the rate of punishment (only 4% (12/268) of the time). Similarly, *Low* enforce roughly the same norms on *High* as they do in One *High* Reveal (in contrast to the norms enforced in *Observed*), and again there is a drop in the punishment rate (only 10% (53/536) of the time compared to 28% in *Observed*).

High types targeting *Low* in Both *High* Reveal continue to enforce a low contribution threshold norm on the extensive margin and a threshold close to the *Low* endowment on the intensive margin (targeting only 5% (25/536) of the time). In this information state *High* types sparingly punish other *High* types (targeting only 3% (8/268) of the time). We estimate a contribution threshold of 25 on the extensive margin, but our model could not converge on the intensive margin.

Finally, Figure 6 shows the likelihood surfaces for models 3-10. In One *High* Reveal we see mostly single-peaked surfaces on the extensive margin, and some flat regions on the intensive margin. The results are noisier in Both *High* Reveal, where sanctioning dropped considerably. The surfaces for *High* are considerably flat across the domains, while the surfaces for *Low* are jagged. Groups in these information states tended to punish less – especially in Both *High* Reveal, and the likelihood surfaces suggest that there was less coordination on contribution norms.

Summary. To summarize our results in this section, groups in *Observed* establish and enforce a "contribute-your-endomment" norm (i.e., contribute the social optimum), replicating the findings from other linear public goods games like Reuben and Riedl (2013) (heterogeneous endowments) and Carpenter and Matthews (2009) (homogeneous endowments). Moreover, there is little disagreement over these norms, and they are enforced by both *High* and *Low* types.

In *Unobserved*, groups in the No *High* and One *High* Reveal states tended to enforce a "contribute-the-*Low*-endowment" norm, and this norm was primarily enforced by *Low* types, while *High* types withdrew from enforcement. When groups had complete information (Both *High* Reveal), there was little enforcement of either *Low* and *High* types.

Overall, these results suggest that private information has an effect on *which* norms emerge and *who* enforces them, relative to complete information, where as expected we see the emergence of efficient contribution norms. In the next and final section we look at *how* norms were enforced across treatments.

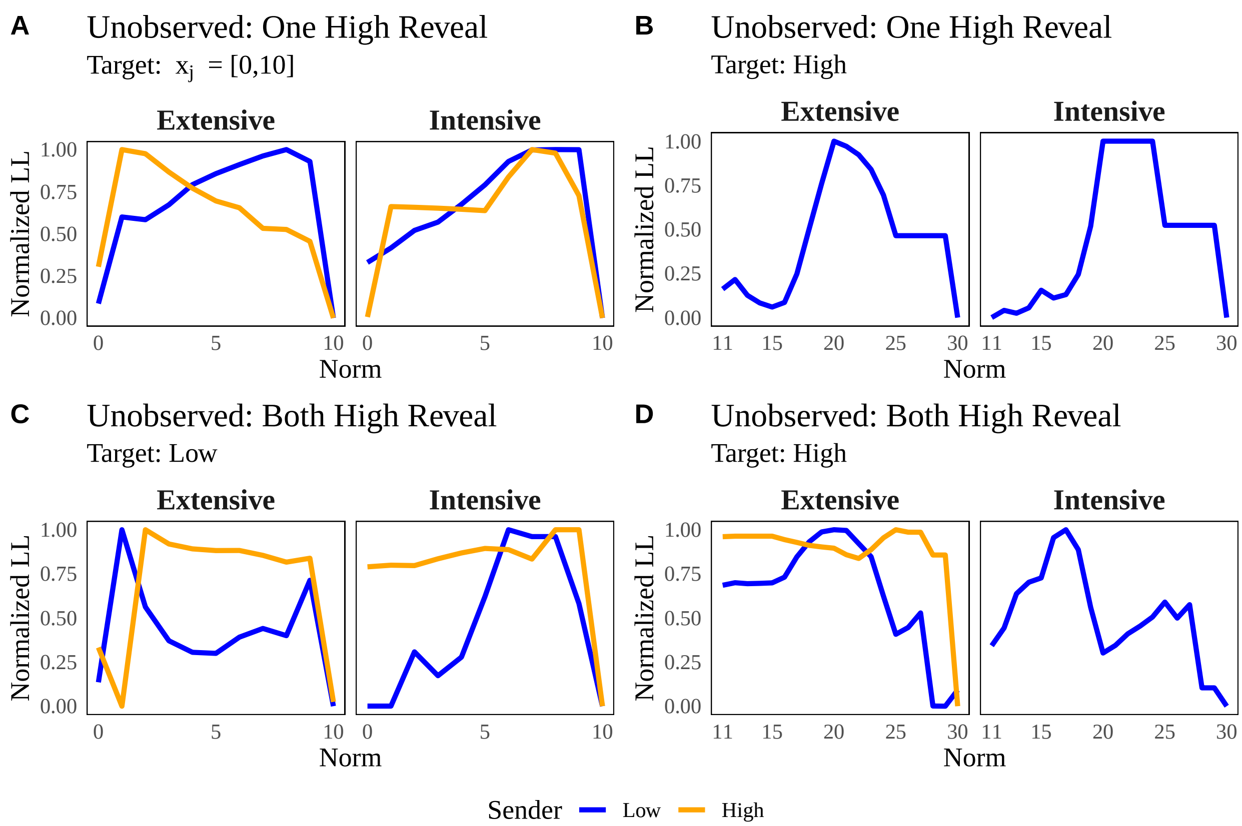


Figure 6: Likelihood surfaces for *Unobserved* in the information states One *High* Reveal and Both *High* Reveal.

4.4 Expected costs and payoffs of noncompliance

In this section we use the estimated contribution norms in each treatment (and in *Unobserved*, each information state) to calculate the expected costs and payoffs of noncompliance with the norms. Since punishment was probabilistic, we trace out the expected costs of noncompliance by combining our estimates for the extensive and intensive margins.[21] We calculated the expected punishment C_{ij} from sender i to target j:

$$C_{ij} = P(s_{ij} > 0|x_j) \times \mathbb{E}[s_{ij}|s_{ij} > 0, x_j] \tag{5}$$

where $P(s_{ij} > 0|x_j)$ is the probability of punishment from i to j and $\mathbb{E}[s_{ij}|s_{ij} > 0]$ is the severity. Both terms are calculated by plugging in j's choice set $[0, e_j]$ into the derivatives of Equations 3 and 4 and evaluated with the estimated parameters at their likelihood-maximizing norms. Equation 4 ensures that predicted sanctions are bounded below at zero. To mimic the punishment technology in our design, we bound predicted sanctions above at 10. We account for the distribution of types within groups (two *Low* and two *High*) when aggregating punishment. For instance, if target j is *High*, they can be targeted by the other *High* and two *Lows* so total expected cost to j is the sum of punishments from two *Low* and one *High*. In *Unobserved* we adjust this calculation to only

[21] In the appendix we plot unconditional and unweighted average punishment for *Low* and *High* across treatments.

include sanctions from *Low* on *High* in One *High* Reveal and Both *High* Reveal, since no norms are estimated for *High* senders in these settings (see Table 3 and Figure 6).

Finally, we use the expected costs to calculate expected payoffs of noncompliance to the norms for subject i:

$$\pi_i = (e_i - x_i) + \alpha \left(x_i + \sum_j x \right) - \mathbb{1}_{x_i < \gamma, x_i < \psi} \sum_j C_{ij} \tag{6}$$

where we assume the other group members $j \neq i$ make socially-optimal contributions. The indicator $\mathbb{1}_{x_i < \gamma, x_i < \psi}$ switches on the punishment $\sum_j C_{ij}$ to subject i when their contribution x_i falls below the norms on the extensive (γ) and intensive (ψ) margins. These contribution norms and expected costs of noncompliance are those we estimate from the data.

Our results are shown in Figure 7. Panels A and B show expected costs of noncompliance (i.e., the expected punishment for each contribution). On the y-axis in red are the theoretical costs of noncompliance to subject i based on the payoff function (Equation 1) and the assumption that all other subjects $-i$ are contributing at the social optimum. For each type, the marginal benefit of noncompliance is equivalent: they can increase their individual payoffs $1 - \alpha = 0.6$ EDs for each ED withheld from the public good. However, the total benefit of noncompliance varies with endowments. When each type contributes their entire endowment they each earn 32 EDs each period. From this reference point, a *High* type can earn 50 EDs by contributing zero (the deterrent penalty is 18 EDs) and a *Low* type can earn 38 EDs by contributing zero (the deterrent penalty is only 6 EDs). To compare our results to theory, we also plot the expected punishment assuming other group members contribute the social optimum and punish at exactly the deterrent level (i.e., the punishment that makes the target indifferent between contributing the social optimum and any other contribution).

In *Observed* we see that the expected costs of noncompliance match the deterrent level for *Low* types but remain below the deterrent level for *High* types. This may simply reflect the higher cost of deterring *High* types relative to *Low* types, but it also helps explain why average contributions remain around 21 among *High* types despite the estimated norm of 30. Importantly, however, the expected costs of noncompliance are substantially higher for *High* types relative to *Low* types at any contribution up to 10. This suggests that groups accounted for their targets' endowment when determining their punishment and attempted to "price discriminate" noncompliance.

We find different patterns of punishments in *Unobserved*. Our estimates of the contribution norms show that groups in private information states (No *High* Reveal and One *High* Reveal) enforce a "contribute-the-*Low*-endowment" norm. The expected cost curves in Panels A and B show that this norm is enforced by punishing contributions of zero as if they came from *High* and contributions of 10 (the *Low* endowment) as if they came from *Low*. The expected costs to contributions of zero are slightly above the deterrent penalty for *High*, but they are about four times greater than the deterrent penalty for *Low*. Underscoring the fact that groups prioritized enforcing a "contribute-the-*Low*-endowment" norm in No *High* and One *High* Reveal, the expected cost of noncompliance to *Low* in Both *High* Reveal flatlines around zero for all contributions. Expected costs to *High* types in *Unobserved* who contributed above 10 are fairly close to the expected costs in *Observed*.

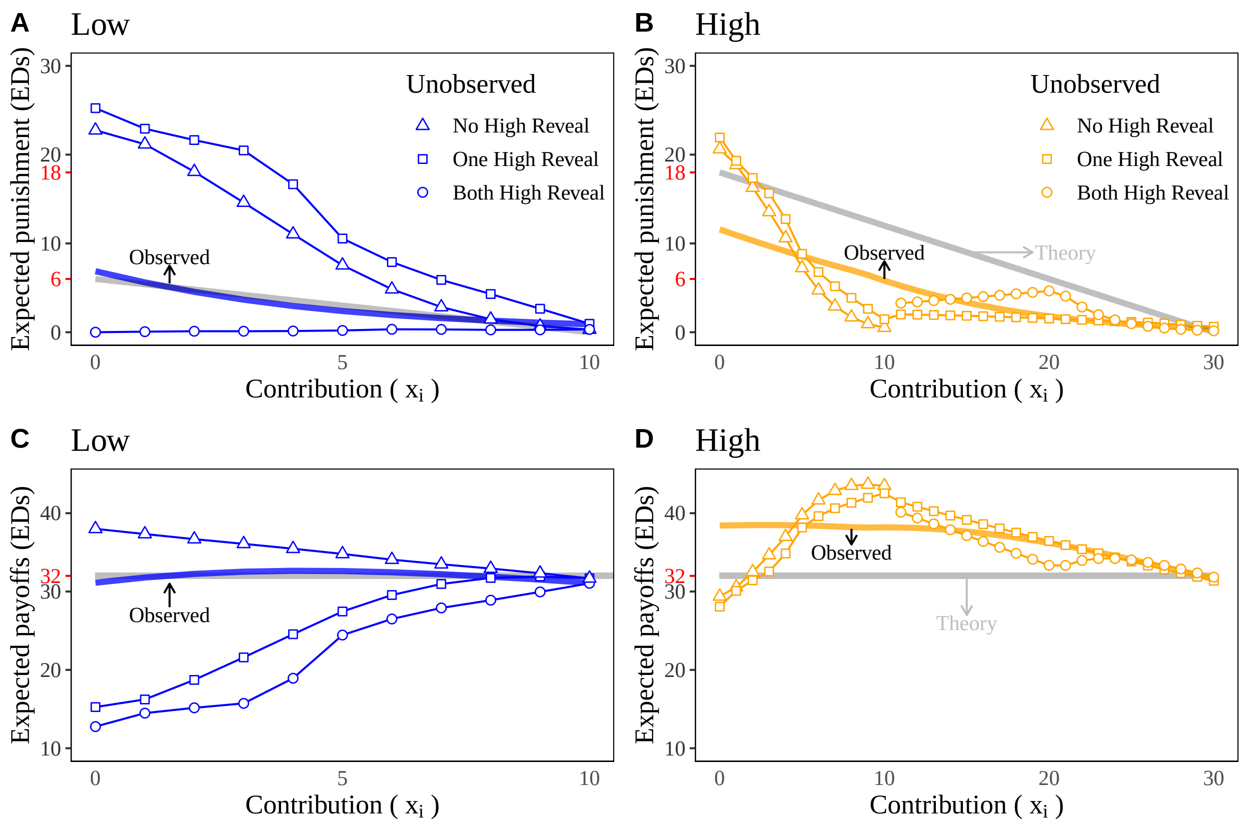


Figure 7: Expected costs (A and B) and payoffs (C and D) to noncompliance with estimated contribution norms. Calculations were made with Equations 5 and 6 and the estimated norms in Tables 2 and 3. To calculate expected payoffs to subject i we varied their contribution while holding the contributions of group members $j \neq i$ at the social optimum. In Panels A and B, theoretical deterrence is drawn in gray, and the penalty for pure freeriding (18 for *High*, 6 for *Low*) is marked in red on the y-axes. In Panels C and D, the socially optimal payoff, 32, is depicted in gray.

Taken together, the results in Figure 7 suggest that subjects (in particular *Low* types) adopted norms in order to balance the benefits of punishing non-cooperative *High* types, while avoiding misguided punishment of cooperative *Low* types. In the absence of complete information (in the No or One *High* Reveal information states), contributions of zero were treated as if they came from *High*, and contributions of 10 as if they came from *Low*. In other words, subjects used punishment to deter pure freeriding, but at the same time created an incentive for *High* to mimic the behavior of a cooperative *Low*.

Given the expected costs of noncompliance to the emergent norms, the expected payoffs depicted in Figure 7 give us a more detailed look at the relative payoffs of the same contributions in different treatments and information states. From the perspective of a *Low* type, the enforced norm is to contribute 10 in both *Observed* and *Unobserved*, and as a consequence, *Low* types contribute a statistically equivalent amount across treatments. In contrast, the enforced norms vary across treatments for *High* types. In *Observed*, *High* types are expected to contribute 30, while in *Unobserved* the emergent norms allow them to mimic the behavior of a cooperative *Low* type (i.e., contribute 10). As a result, *High* types who contributed 10 are better off in *Unobserved* than in *Observed*, while expected payoffs to contributions above 10 are roughly the same across treatments. Alternatively, *Low* types who contribute 10 had the same expected payoffs as their counterparts in *Observed*, regardless of the information state.

5 Discussion and concluding remarks

We study how groups in a public goods game with heterogeneous endowments (*Low* and *High* types) use peer enforcement to adapt to private information (subjects observe contributions but not endowments). In the absence of complete information, there is a risk that sanctions may mistakenly target cooperative behavior. We show that groups manage this risk by establishing contribution norms and enforcement strategies that: (i) prevent *High* types from freeriding; (ii) reward *Low* types for cooperating; but as a consequence (iii) enable *High* types to hide behind "small endowments", similar to how proposers in bargaining games with private information hide behind "small cakes" (Güth et al., 1996; Mitzkewitz and Nagel, 1993).[22]

These norms and enforcement strategies prevented cooperation from unraveling, but they also kept groups stuck in an inferior equilibrium relative to complete information. The benefits to information in social dilemmas thus appear to be twofold. First, transparency allows groups to map behavior to capacities (e.g., contributions to endowments), paving the way for efficient norms (i.e., norms that maximize cooperation) to emerge. Second, our estimates of the expected costs of noncompliance show that transparency allows groups to price discriminate when enforcing norm compliance by adjusting punishment to account for heterogeneity. When agents have private information and exploit it, we still see cooperation-enhancing norms emerge, but first-degree price discrimination (i.e., the cost of noncompliance depends of your endowment) is no longer possible, and thus the norms that maximize cooperation cannot emerge purely from the disincentives to non-cooperation created by peer enforcement.

There are a number of ways to extend our paper. For starters, future work can explore norm emergence in larger groups, groups with different distributions of endowments, or both. In addition, power asymmetries may play an important role. In our design we restricted *Low* and *High* types to the same enforcement budget each round, so *ex ante* neither type had an out-sized influence on which norms would emerge. In reality, agents with more resources often have more power to influence outcomes at the macro-level (e.g. economic growth, Acemoglu et al., 2005) and the micro-level (e.g. the formation and enforcement of property rights, Waichman, 2020; Jayadev and Bowles, 2006), and this could influence the norms and enforcement strategies that emerge in our design.

Future work can also study whether group endogeneity can solve the problem of establishing

[22] Similar behavior has been found in dictator games. Dana et al. (2007) show that dictators exploit opportunities for "moral wiggle room". When dictators are given the opportunity to obscure the relationship between their behavior and the resulting outcomes they take advantage and act more in their self-interest.

socially-optimal norms under private information. Our results show that socially-optimal norms do not emerge in these exogenous groups, and the norms that do emerge mitigate the costs of private information through deliberate peer enforcement. But in reality, social organization is endogenous to some degree: individuals flock together either through self-sorting, by the hand of some authority (e.g., the manager of a firm), or some combination thereof.[23]

One reason why endogeneity might matter is because individuals bring their preferences into groups. Fehr and Williams (2013) show that self-sorting can produce efficient norms in social dilemmas, even in the absence of punishment, because some individuals are conditional cooperators by nature. Kimbrough and Vostroknutov (2016) make a similar point; they suggest that pro-social behavior can be explained by the fact that some subjects come into experiments with preferences to obey norms, even arbitrary ones.[24] Kimbrough and Vostroknutov (2016) identify subjects who are inherent rule-followers and show that they can sustain cooperation in a public goods game without punishment when paired with other rule-followers, but not when they are paired with inherent rule-breakers.

Moreover, there is evidence that inherent preferences to follow rules can produce cooperative behavior even in settings with private information. For instance, most cultures teach people not to lie, and this norm seems to reduce lying to some degree. Abeler et al. (2019) find that in truth-telling experiments, where lying cannot be directly punished, subjects around the world lie much less than their monetary incentives alone would predict. Abeler et al. (2019) argue this is because people like to be seen as honest *and* because they have a preference for being honest.[25] One way to think about these subjects is they are disciplined by the internalized norm "don't lie". Speculating on this result, the internalization of certain norms and the costs of noncompliance (e.g., guilt or some other psychological cost)[26] could lead to preferences for norm compliance, even when private information can be exploited. If such agents were put in a social dilemma with private information and given a mechanism to flock together, it is possible they could establish socially-optimal norms, and perhaps even produce spillover benefits to agents without such preferences (as in Fehr and Williams, 2013).

[23] There is some evidence from industry that endogenous groups allow for private information without compromising cooperation. For instance, software engineers and sales representatives at Google are grouped by managers into teams that rely on the cooperation of team members to produce output. Google studied these teams and found that concerns about private information are in fact quite low; the highest-performing teams do not promote cooperation by establishing rules that emphasize monitoring and enforcement (Duhigg, 2016).

[24] Kimbrough and Vostroknutov (2016) point out that subjects often bring well-established norms into the lab (e.g. "do not lie", "split a surplus fifty-fifty") that regulate their behavior.

[25] A desire to be seen as honest (or fair) is also an argument used to explain behavior in other settings where a norm cannot be enforced like dictator games (Ockenfels and Werner, 2012; Andreoni and Bernheim, 2009) and charitable giving experiments (Grossman, 2015).

[26] Bucciol and Piovesan (2011) run a truth-telling task with children age 5 to 15 and find that a) children lie but less than they could, and b) children lie even less when the experimenter makes a normative appeal to tell the truth. Talwar et al. (2015) also show that normative appeals reduce lying among children in a non-incentivized task. Interestingly, the authors also show that "expected punishments" (a child is told if they lie "you will be in trouble", but no punishment is actually carried out) crowd-out the effect of normative appeals and lead to an increase in lying.

References

Abbink, K., L. Gangadharan, T. Handfield, and J. Thrasher (2017). Peer punishment promotes enforcement of bad social norms. *Nature communications 8* (1), 1–8.

Abeler, J., D. Nosenzo, and C. Raymond (2019). Preferences for truth-telling. *Econometrica 87* (4), 1115–1153.

Acemoglu, D., S. Johnson, and J. A. Robinson (2005). Institutions as a fundamental cause of long-run growth. *Handbook of Economic Growth 1*, 385–472.

Ambrus, A. and B. Greiner (2012). Imperfect public monitoring with costly punishment: An experimental study. *The American Economic Review 102* (7), 3317–3332.

Andreoni, J. and B. D. Bernheim (2009). Social image and the 50–50 norm: A theoretical and experimental analysis of audience effects. *Econometrica 77* (5), 1607–1636.

Arrow, K. (1970). Political and economic evaluation of social effects and externalities. In *The analysis of public output*, pp. 1–30. NBER.

Bacchelli, A. and C. Bird (2013). Expectations, outcomes, and challenges of modern code review. In *2013 35th International Conference on Software Engineering (ICSE)*, pp. 712–721. IEEE.

Bochet, O., T. Page, and L. Putterman (2006). Communication and punishment in voluntary contribution experiments. *Journal of Economic Behavior & Organization 60* (1), 11–26.

Bornstein, G. and O. Weisel (2010). Punishment, cooperation, and cheater detection in “noisy” social exchange. *Games 1*, 18–33.

Bowles, S. and S.-H. Hwang (2008). Social preferences and public economics: Mechanism design when social preferences depend on incentives. *Journal of public economics 92* (8-9), 1811–1820.

Bucciol, A. and M. Piovesan (2011). Luck or cheating? A field experiment on honesty with children. *Journal of Economic Psychology 32* (1), 73–78.

Buckley, E. and R. Croson (2006). Income and wealth heterogeneity in the voluntary provision of linear public goods. *Journal of Public Economics 90*, 935–955.

Cappelen, A. W., A. D. Hole, E. Ø. Sørensen, and B. Tungodden (2007). The pluralism of fairness ideals: An experimental approach. *American Economic Review 97* (3), 818–827.

Carpenter, J. and P. H. Matthews (2009). What norms trigger punishment? *Experimental Economics 12* (3), 272–288.

Carpenter, J. P. (2004). When in Rome: Conformity and the provision of public goods. *The Journal of Socio-Economics 33* (4), 395–408.

Cason, T. N. and L. Gangadharan (2015). Promoting cooperation in nonlinear social dilemmas through peer punishment. *Experimental Economics 18* (1), 66–88.

Chaudhuri, A. (2011). Sustaining cooperation in laboratory public goods experiments: a selective survey of the literature. *Experimental Economics 14* (1), 47–83.

Cherry, T. L., S. Kroll, and J. Shogren (2005). The impact of endowment heterogeneity and origin on public good contributions: evidence from the lab. *Journal of Economics Behavior and Organization 57*, 357–365.

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

Cressman, R. and Y. Tao (2014). The replicator equation and other game dynamics. *Proceedings of the National Academy of Sciences 111* (Supplement 3), 10810–10817.

Dana, J., R. A. Weber, and J. X. Kuang (2007). Exploiting moral wiggle room: experiments demonstrating an illusory preference for fairness. *Economic Theory 33* (1), 67–80.

De Geest, L. R. and D. C. Kingsley (2019). Endowment heterogeneity, incomplete information & institutional choice in public good experiments. *Journal of Behavioral and Experimental Economics*, 101478.

De Geest, L. R. and J. K. Stranlund (2019). Defending public goods and common-pool resources. *Journal of Behavioral and Experimental Economics 79*, 143–154.

Duhigg, C. (2016). What google learned from its quest to build the perfect team. *The New York Times Magazine 26* (2016), 2016.

Ertan, A., T. Page, and L. Putterman (2009). Who to punish? Individual decisions and majority rule in mitigating the free rider problem. *European Economic Review 53* (5), 495–511.

Fehr, E. and S. Gächter (2000). Cooperation and punishment in public goods experiments. *American Economic Review 90* (4), 980–994.

Fehr, E. and I. Schurtenberger (2018). Normative foundations of human cooperation. *Nature Human Behaviour 2* (7), 458–468.

Fehr, E. and T. Williams (2013). Endogenous emergence of institutions to sustain cooperation. Technical report, mimeo.

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

Gächter, S., E. Renner, and M. Sefton (2008). The long-run benefits of punishment. *Science 322* (5907), 1510–1510.

Gintis, H., E. A. Smith, and S. Bowles (2001). Costly signaling and cooperation. *Journal of Theoretical Biology 213*(1), 103–119.

Grechenig, K., A. Nicklisch, and C. Thöni (2010). Punishment despite reasonable doubt? A public goods experiment with sanctions under uncertainty. *Journal of Empirical Legal Studies 7*(4), 847–867.

Greiner, B. (2015). Subject pool recruitment procedures: organizing experiments with ORSEE. *Journal of the Economic Science Association 1*(1), 114–125.

Grossman, Z. (2015). Self-signaling and social-signaling in giving. *Journal of Economic Behavior & Organization 117*, 26–39.

Güth, W., S. Huck, and P. Ockenfels (1996). Two-level ultimatum bargaining with incomplete information: An experimental study. *The Economic Journal 106*(436), 593–604.

Hargreaves Heap, S. P., A. Ramalingam, and B. V. Stoddard (2016). Endowment inequality in public goods games: A re-examination. *Economics Letters 146*, 4–7.

Hofmeyr, A., J. Burns, and M. Visser (2007). Income inequality, reciprocity and public good provision: An experimental analysis. *South African Journal of Economics 75*(3), 508–520.

Jayadev, A. and S. Bowles (2006). Guard labor. *Journal of Development Economics 79*(2), 328–348.

Kimbrough, E. O. and A. Vostroknutov (2016). Norms make preferences social. *Journal of the European Economic Association 14*, 608–638.

Kimbrough, E. O. and A. Vostroknutov (2018). A portable method of eliciting respect for social norms. *Economics Letters 168*, 147–150.

Kingsley, D. C. (2016). Endowment heterogeneity and peer punishment in a public good experiment: Cooperation and normative conflict. *Journal of Behavioral and Experimental Economics 60*, 49–61.

Krupka, E. L. and R. A. Weber (2013). Identifying social norms using coordination games: Why does dictator game sharing vary? *Journal of the European Economic Association 11*(3), 495–524.

Miller, J. H. and J. Andreoni (1991). Can evolutionary dynamics explain free riding in experiments? *Economics Letters 36*(1), 9–15.

Mitzkewitz, M. and R. Nagel (1993). Experimental results on ultimatum games with incomplete information. *International Journal of Game Theory 22*(2), 171–198.

Nicklisch, A., K. Grechenig, and C. Thöni (2016). Information-sensitive Leviathans. *Journal of Public Economics 144*, 1–13.

Nicklisch, A. and I. Wolff (2011). Cooperation norms in multiple-stage punishment. *Journal of Public Economic Theory 13* (5), 791–827.

Nikiforakis, N. (2008). Punishment and counter-punishment in public good games: Can we really govern ourselves? *Journal of Public Economics 92* (1-2), 91–112.

Nikiforakis, N. and H.-T. Normann (2008). A comparative statics analysis of punishment in public-good experiments. *Experimental Economics 11* (4), 358–369.

Nikiforakis, N., C. N. Noussair, and T. Wilkening (2012). Normative conflict and feuds: The limits of self-enforcement. *Journal of Public Economics 96* (9-10), 797–807.

Ockenfels, A. and P. Werner (2012). 'Hiding behind a small cake' in a newspaper dictator game. *Journal of Economic Behavior & Organization 82* (1), 82–85.

Reuben, E. and A. Riedl (2013). Enforcement of contribution norms in public good games with heterogeneous populations. *Games and Economic Behavior 77* (1), 122–137.

Robbett, A. (2016). Sustaining cooperation in heterogeneous groups. *Journal of Economic Behavior and Organization 132*, 121–138.

Talwar, V., C. Arruda, and S. Yachison (2015). The effects of punishment and appeals for honesty on children's truth-telling behavior. *Journal of Experimental Child Psychology 130*, 209–217.

Visser, M. and B. J. (2015). Inequality, social sanctions and cooperation within south african fishing communities. *Journal of Economic Behavior and Organization 118*, 95–109.

Vostroknutov, A. (2020). Social norms in experimental economics: Towards a unified theory of normative decision making. *Analyse & Kritik 42* (1), 3–40.

Waichman, I. (2020). Linking wealth and punishment effectieness: Punishment and cooperation under congruent heterogeneities. *Economic Inquiry 58* (1), 86–103.

Weng, Q. and F. Carlsson (2015). Cooperation in teams: The role of identity, punishment, and endowment distribution. *Journal of Public Economics 126*, 25–38.

Zelmer, J. (2003). Linear public good experiments: A meta-analysis. *Experimental Economics 6* (3), 299–310.

A Coefficient estimates of models at likelihood-maximizing norms

A.1 Extensive margin of punishment

Table 4: Estimated probability of punishment. Each model shows estimates for Equation 3 at the likelihood-maximizing norms contribution norms. Models are organized by Sender $\rightarrow$ Target. For instance in *Observed*, $L \rightarrow L$ means "Low targeting Low"; in *Unobserved* $L \rightarrow [0, 10]$ means "Low targeting contributions between zero and ten".

	Observed				No High Reveal		One High Reveal			Both High Reveal			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)
	$L \rightarrow L$	$L \rightarrow H$	$H \rightarrow L$	$H \rightarrow H$	$L \rightarrow [0,10]$	$H \rightarrow [0,10]$	$L \rightarrow [0,10]$	$L \rightarrow H$	$H \rightarrow [0,10]$	$L \rightarrow L$	$L \rightarrow H$	$H \rightarrow L$	$H \rightarrow H$
	$\hat{\gamma} = 9$	$\hat{\gamma} = 29$	$\hat{\gamma} = 9$	$\hat{\gamma} = 29$	$\hat{\gamma} = 9$	$\hat{\gamma} = 1$	$\hat{\gamma} = 8$	$\hat{\gamma} = 20$	$\hat{\gamma} = 1$	$\hat{\gamma} = 1$	$\hat{\gamma} = 20$	$\hat{\gamma} = 2$	$\hat{\gamma} = 25$
Target Contribution	-0.462***	-0.163***	-0.131	-0.030	-0.296***	-0.892***	-0.220***	-0.172	-3.075***	9.385***	0.089	-8.735***	-0.885
	(0.13)	(0.05)	(0.09)	(0.04)	(0.08)	(0.33)	(0.03)	(0.12)	(0.68)	(2.73)	(0.10)	(1.97)	(0.57)
Contribution	0.010	0.053***	-0.017	0.012	0.012	-0.001	0.071***	0.137**	0.023	-0.212***	0.282**	0.018	-0.148
	(0.06)	(0.02)	(0.01)	(0.01)	(0.05)	(0.07)	(0.02)	(0.05)	(0.04)	(0.08)	(0.14)	(0.10)	(0.42)
Average Contribution	0.159**	0.130**	0.171**	0.057	0.225	0.257	0.172***	0.021	0.051	0.819***	0.502	1.758***	2.168
	(0.07)	(0.06)	(0.08)	(0.06)	(0.14)	(0.16)	(0.05)	(0.12)	(0.20)	(0.24)	(0.35)	(0.31)	(1.83)
Lagged sanctions	0.024	0.010	0.018**	0.012	0.017	0.041***	0.019*	0.073***	0.078*	0.110**	0.040***	0.019	-0.020
	(0.03)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)	(0.04)	(0.04)	(0.01)	(0.02)	(0.04)
Period	-0.022**	-0.012	-0.007	-0.012	-0.004	-0.011	-0.002	-0.028	-0.011	0.045***	0.071***	0.055	0.091***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)	(0.01)	(0.01)	(0.01)	(0.04)	(0.02)
Deviation	0.666	0.800	0.525	1.829	1.175***	0.804**	2.810***	1.483***	2.975***	-10.271***	-0.456	11.685***	0.000
	(0.96)	(1.22)	(0.55)	(1.44)	(0.38)	(0.39)	(0.58)	(0.35)	(0.81)	(3.19)	(1.06)	(2.56)	(.)
Average contribution X Deviation	-0.163***	-0.211***	-0.117***	-0.175**	-0.085**	-0.018	-0.223***	-0.125***	0.012	-0.026	-0.013	-0.252***	0.000
	(0.06)	(0.06)	(0.04)	(0.08)	(0.04)	(0.03)	(0.04)	(0.04)	(0.03)	(0.03)	(0.06)	(0.05)	(.)
Constant	-0.223	0.328	-1.887	-0.798	-1.029**	-1.358**	-1.518***	-0.028	-0.181	-14.686	-13.261*	-6.562***	-17.387**
	(0.82)	(0.79)	(1.18)	(0.89)	(0.48)	(0.62)	(0.53)	(1.07)	(1.96)	(.)	(7.71)	(1.19)	(7.63)
N (extensive)	882	1764	1764	882	1608	1608	360	180	450	264	528	528	128
N (intensive)	45	234	240	158	262	131	107	51	66	11	53	24	8

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

A.2 Intensive margin of punishment

Table 5: Estimated probability of punishment. Each model shows estimates for Equation 3 at the likelihood-maximizing norms contribution norms. For models with multiple candidate norms (e.g., (8) and (9)) we estimated the coefficients at the smallest norm. Models are organized by Sender $\rightarrow$ Target. For instance in *Observed*, $L \rightarrow L$ means "Low targeting Low"; in *Unobserved* $L \rightarrow [0, 10]$ means "Low targeting contributions between zero and ten".

	Observed				No High Reveal		One High Reveal			Both High Reveal		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	$L \rightarrow L$	$L \rightarrow H$	$H \rightarrow L$	$H \rightarrow H$	$L \rightarrow [0,10]$	$H \rightarrow [0,10]$	$L \rightarrow [0,10]$	$L \rightarrow H$	$H \rightarrow [0,10]$	$L \rightarrow L$	$L \rightarrow H$	$H \rightarrow L$
	$\hat{\psi} = 9$	$\hat{\psi} = 29$	$\hat{\psi} = 9$	$\hat{\psi} = 28$	$\hat{\psi} = 9$	$\hat{\psi} = 9$	$\hat{\psi} = 8$	$\hat{\psi} = [20-24]$	$\hat{\psi} = 7$	$\hat{\psi} = 6$	$\hat{\psi} = 17$	$\hat{\psi} = [8-9]$
Target Contribution	-0.108	-0.021**	-0.024	-0.089***	-0.192***	-0.098	-0.094	0.069***	-0.161**	-0.220***	-0.394	-0.199***
	(0.23)	(0.01)	(0.22)	(0.03)	(0.07)	(0.07)	(0.08)	(0.02)	(0.07)	(0.00)	(0.44)	(0.05)
Contribution	0.050*	0.085	0.011	-0.037	-0.077	0.008	0.030	0.022	-0.011	-0.045***	-0.379**	-0.049
	(0.03)	(0.20)	(0.05)	(0.06)	(0.06)	(0.07)	(0.06)	(0.02)	(0.07)	(0.00)	(0.16)	(0.19)
Average Contribution	0.014	0.027	0.025	0.209	0.167	0.043	-0.004	0.024	0.116	0.047***	0.494	0.201
	(0.04)	(0.07)	(0.13)	(0.19)	(0.11)	(0.08)	(0.30)	(0.03)	(0.25)	(0.00)	(0.35)	(0.23)
Lagged sanctions	-0.021	-0.017	0.002	-0.019	0.004	0.004	0.019	0.031***	-0.001	-0.121***	-0.008	0.063
	(0.03)	(0.03)	(0.04)	(0.02)	(0.07)	(0.01)	(0.04)	(0.01)	(0.03)	(0.00)	(0.03)	(0.05)
Period	-0.014**	0.003	-0.001	0.001	0.005	0.013**	0.016*	0.021***	0.013	-0.115***	-0.006	0.056***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.00)	(0.04)	(0.00)	(0.02)	(0.01)
Deviation	3.235	-0.488	1.533	1.230*	1.474	-0.020	-0.639*	0.575***	0.584	4.313***	1.325	0.408
	(3.65)	(0.47)	(2.53)	(0.71)	(1.25)	(2.35)	(0.35)	(0.12)	(1.60)	(0.01)	(1.85)	(1.50)
Average contribution X Deviation	-0.145	-0.006	-0.073	-0.047	-0.118	0.039	0.089	-0.057***	-0.022	-0.132***	-0.056	-0.073
	(0.12)	(0.03)	(0.21)	(0.04)	(0.08)	(0.31)	(0.06)	(0.01)	(0.15)	(0.00)	(0.08)	(0.10)
Constant	0.421	0.031	0.300	0.375	1.023**	0.683	0.427	-1.502***	0.431	1.006***	3.149	-0.297
	(0.29)	(0.48)	(0.52)	(0.84)	(0.46)	(0.55)	(2.42)	(0.27)	(2.59)	(0.01)	(2.92)	(1.77)
N (extensive)	882	1764	1764	882	1608	1608	360	180	450	264	528	528
N (intensive)	45	234	240	158	262	131	107	51	66	11	53	24

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

B An evolutionary model of punishment with private information

Our empirical results suggest that the contribution norms that emerge under incomplete information mitigate but do not eliminate non-cooperative behavior from advantaged agents. That is, *High* types in *Unobserved* cannot get away with contributing nothing, but they can get away with contributing 10 and hiding behind a small endowment.

In our results we argue this is because groups wanted to ensure some cooperation from *High* types while avoiding misguided punishment of *Low* types. But could the same or an even better outcome obtained using a different rule, like "punish all contributions of 10 as if they came from *High*"? We explore this idea with a simple evolutionary model based on our design.

Suppose two agents, one *Low* and one *High*, meet to play a public goods game with the same payoff function as our experiment, except we now set the MPCR to 0.8 (to ensure $\frac{1}{n} < \alpha < 1$) and restrict *Low* to two strategies (contribute 0 or 10) and *High* to three strategies (contribute 0, 10, or 30). As usual, the one-shot Nash equilibrium and subgame perfect equilibrium with a known end period is mutual defection (both contribute 0). Since our main interest is understanding the motives behind the emergence of the punishment rules in Figure 7, we will consider what happens when strategies in this game face different punishment rules imposed top-down by a social planner and evolve according to a standard replicator dynamic used in other models of public goods games (e.g. Cressman and Tao, 2014; Carpenter, 2004; Gintis et al., 2001; Miller and Andreoni, 1991).[27]

Assuming a large, well-mixed population of a fixed size, the replicator dynamic describes how the proportion of the population playing a given strategy evolves from one period to the next based on the fitness or payoffs to that strategy. Consider a *Low* type who contributes 10. In a population of size N the fraction of *Low* types who contribute 10 is $L_{10} = \frac{N_{L10}}{N}$. The fitness to L_{10} is then the sum of payoffs to playing 10 weighted by the share of each strategy in the population:

$$f_{L_{10}} = L_0\pi(10,0) + H_0\pi(10,0) + H_{10}\pi(10,10) + H_{30}\pi(10,30) \tag{B.1}$$

where $H_{10}\pi(10,10)$ is the weighted payoff to *Low* when they contribute 10 and *High* contributes 10, and so on. Average population fitness is just the sum of these fitnesses weighted by the proportion of agents playing any of the five strategies:

$$\bar{f} = L_{10}f_{L_{10}} + L_0 f_{L_0} + H_{10}f_{H_{10}} + H_{30}f_{H_{30}} + H_0 f_{H_0} \tag{B.2}$$

so the replicator dynamic for any strategy, for example *Low* playing 10, is then

$$\frac{dL_{10}}{dt} = \dot{L}_{10} = L_{10}(f_{L_{10}} - \bar{f}) \tag{B.3}$$

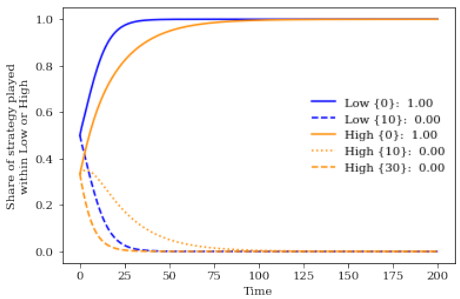
with $\bar{f}$ or average fitness coming from Equation B.1. Equation B.3 simply says that the share of *Low* types contributing 10 will increase over time when the fitness of contributing 10 is greater than

[27] Carpenter (2004) points out that replicator dynamics are a convenient way to mimic the learning process of groups in experiments.

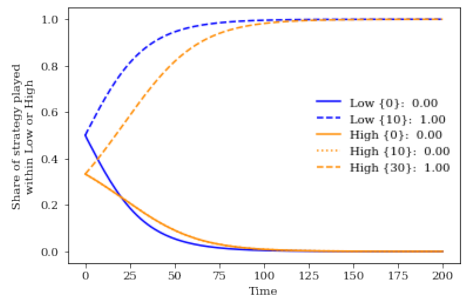
the average fitness, and will decrease if the opposite is the case.

Suppose punishments are meted out by a social planner at zero cost. Since we have an equal population of *Low* and *High* types, we assume half the population of strategies are split among *Low*'s two strategies and the other half are split among *High*'s three strategies.

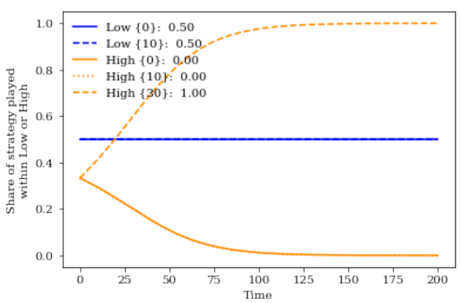
Figure B.1 shows different simulations of this model under different punishment rules. *Low* is in blue and *High* is in orange, and the legend displays the population share of a strategy in the final time step. When there is no punishment, freeriding sweeps through the population (Panel A). But if the social planner can observe the endowment of each type and apply exactly deterrent penalties according to the target's endowment (Panel B), cooperation emerges and both types contribute their full endowments.



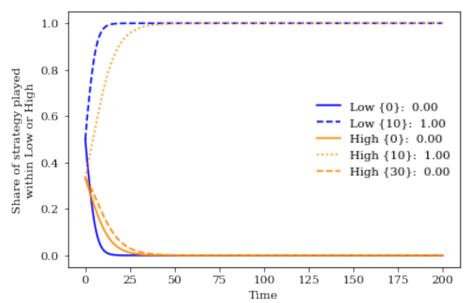
(a) No punishment.



(b) *Observed*: Optimal punishment.



(c) *Unobserved*: punish all 0 and 10 as *High*.



(d) *Unobserved*: punish all 0 as *High* and 10 as *Low*.

Figure B.1: Replicator dynamics for a linear public good game with punishment, two endowment types and complete or incomplete information.

Now suppose the planner cannot observe endowments and must choose a punishment rule subject to this constraint. One enforcement is “coercive” in that it secures full cooperation from *High* by punishing all instances of 0 and 10 as if they came from *High*. Panel C shows the outcome of this scenario. This strategy does indeed secure full cooperation from *High*, but it also reduces cooperation from *Low* by half and replaces it with *Low* freeriding, since both strategies earn the same payoff net of their respective punishments when *High* is fully cooperative.

Alternatively, the planner could instead adopt a more “forgiving” punishment rule similar to what we see in our results: punish all contributions of zero as if they came from *High*, and punish no contributions of 10, in order to avoid punishing cooperative *Low* types. This leads to Panel D.

Similar to our experimental results, we see full cooperation from *Low*, and partial cooperation by *High*.

Interestingly, this "forgiving" punishment rule generates *lower* total contributions ($10 + 10 + 10 + 10 = 40$) than the more "coercive" punishment rule that generates full contributions from *High* ($30 + 30 + 0 + 0 = 60$). It is therefore plausible that groups in our experiment evolved punishment rules that reflected a desire to protect *Low* rather than attack *High*, even if it means tolerating a certain level of non-cooperation, and a lower value of the public good.

C Unconditional average punishment

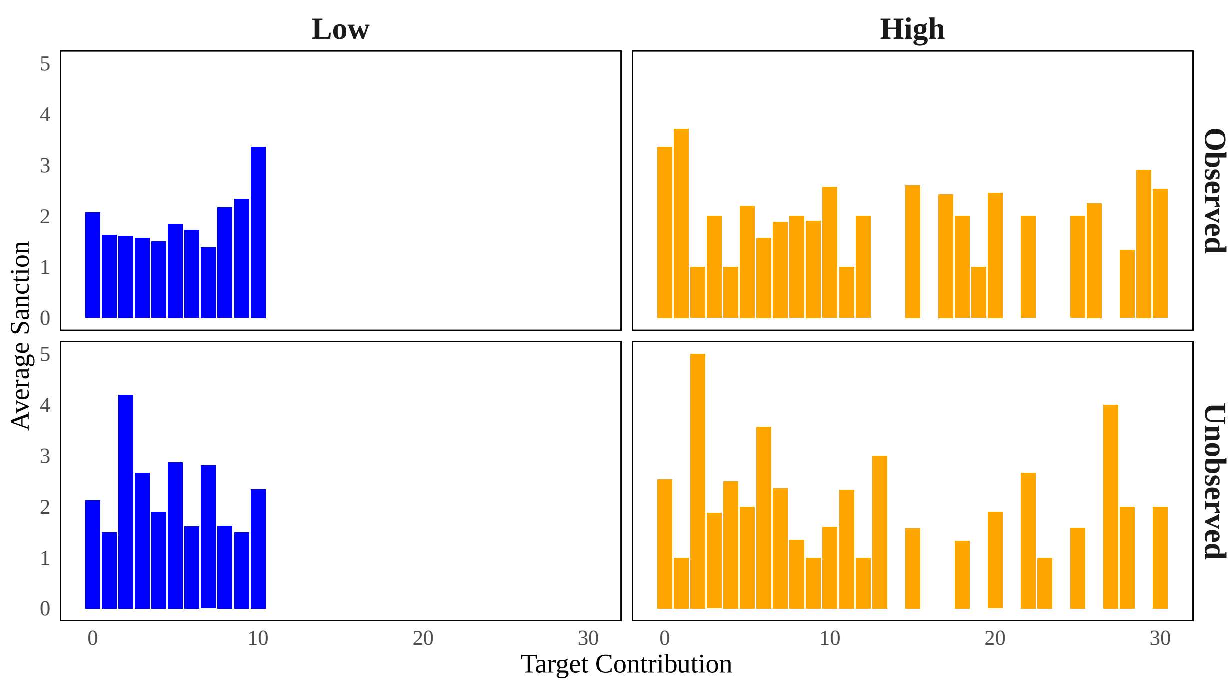


Figure C.1: Unconditional average punishment by endowment type and treatment.

D Experiment instructions

1 Welcome to the Experiment

Thank you for participating in our decision making experiment. The experiment consists of **50 periods**. In each period you will have an opportunity to earn money, which is in addition to the $5 guaranteed for your participation in the experiment. Your earnings each period will depend on your decisions and the decisions of other participants.

Please read the following instructions carefully. Everyone must correctly answer the comprehension questions at the end before we can begin.

During the experiment you are not allowed to communicate with other participants. If you have a question please raise your hand.

During the experiment your earnings will be calculated in *Experimental Dollars* (*EDs* for short). You can earn *EDs* every period. At the end of the experiment, your total earnings in *EDs* will be converted to U.S. dollars at the following rate:

$$100EDs = \$1$$

At the end of the experiment your total earnings (including the $5 participation payment) will be paid to you, privately and anonymously, in cash.

In the experiment, each participant is randomly assigned to a group of 4. This means that you are in a group with 3 other participants. You will be part of the **same group** throughout the entire experiment. However, at no point will the members of your group be revealed. All of the decisions you make within the experiment are anonymous and will be kept confidential.

In every period, each group member, yourself included, will be given an endowment of EDs. Two (2) members of the group will receive 30 EDs and two (2) members of the group will receive 10 EDs. This initial allocation of EDs is random and will be maintained throughout the experiment. Whatever your endowment is in Period 1 will remain your endowment for the entire experiment.

Each period consists of two stages. We will discuss both stages in detail, along with examples, and ask you to complete comprehension questions before starting the experiment.

Stage 1

Each of you will independently and anonymously decide how many of your EDs to allocate to the group account. You can allocate any integer between 0 and your endowment to the group account. Your remaining EDs will automatically be allocated to your private account. Your earnings depend on the number of EDs in your private account and the *total* number of EDs in the group account.

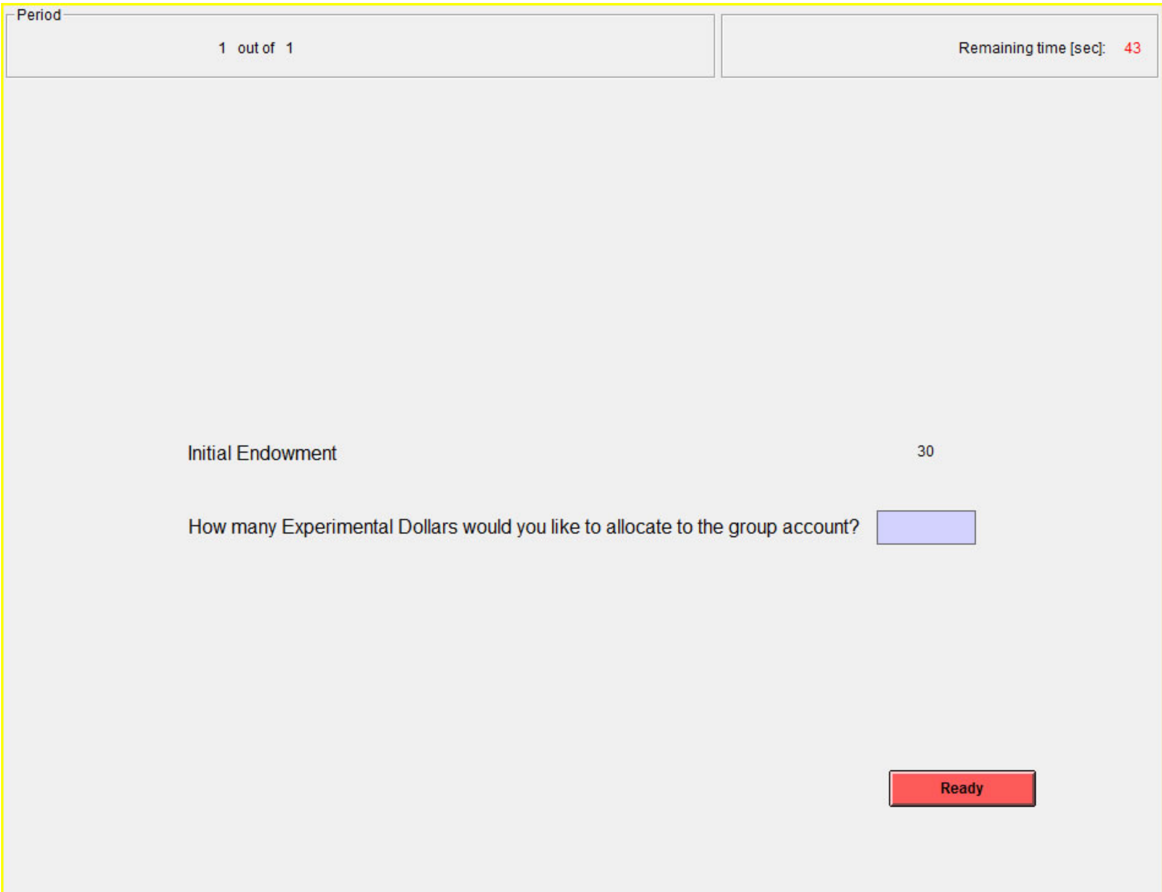


Figure 1: Example of allocation decision screen (assumes a 30 ED endowment)

How are period earnings calculated?

The earnings from your private account equal the number of EDs in your private account. Your private account earnings do not depend on the decisions of other group members. You simply keep all EDs that you choose *not* to allocate to the group account.

Your Private Account Earnings = (Your Endowment) - (Your allocation to group account)

Your earnings from the group account equal 0.4 times the *total* number of EDs allocated to the group account. Thus, your group account earnings depend, in part, on the decisions of other group members.

Your Group Account Earnings = 0.4*(the total number of EDs allocated to the group account)

Your period earnings are the sum of your private account earnings and your group account earnings.

Your Period Earnings = Your Private Account Earnings + Your Group Account Earnings

After Period 1 you will be presented with the history of your choices from previous periods. This information includes the information above and your total earnings up to this point in the experiment. Your total earnings are the sum of your earnings from each period of the experiment.

Your Total Earnings = Sum of your Private Earnings each Period

Example 1

The example assumes the following:

	Endowment	Allocation to Group Account
You	30 EDs	15 EDs
Member A	30 EDs	30 EDs
Member B	10 EDs	10 EDs
Member C	10 EDs	0 EDs

The total number of EDs in the group account = 15+30+10+0 = 55 EDs, so each group member earns = 0.4*55 = 22 EDs from the group account.

What are your period earnings in this example?
You have a 30 ED endowment and you allocated 15 EDs:

Your period earnings = private account earnings + group account earnings
= (your endowment – your allocation) + (0.4*total group allocation)
= (30 – 15) + 0.4*55
= 15 + 22 = 37 EDs

What are the period earnings of Member A in this example?
Group member A has a 30 ED endowment and allocated 30 EDs:

Their period earnings = (their endowment – their allocation) + (0.4*total group allocation)
= (30 – 30) + 0.4*55
= 0 + 22 = 22 EDs

What are the period earnings of Member B in this example?
Group member B has a 10 ED endowment and allocated 10 EDs:

Their period earnings = (their endowment – their allocation) + (0.4*total group allocation)
= (10 – 10) + 0.4*55
= 0 + 22 = 22 EDs

What are the period earnings of Member C in this example?
Group member C has a 10 ED endowment and allocated 0 EDs:

Their period earnings = (their endowment – their allocation) + (0.4*total group allocation)
= (10 – 0) + 0.4*55
= 10 + 22 = 32 EDs

Note that, regardless of your endowment, for each ED you allocate to the group account, your earnings from the group account ***increase*** by 0.4*1 = 0.4 EDs and your earnings from your private account ***decrease*** by 1 ED.

However, for each ED you allocate to the group account, the earnings of each of the other 3 members of your group ***increase*** by 0.4 EDs. Therefore, for each ED you allocate to the group account the total group earnings ***increase*** by 0.4*3 = 1.2 EDs.

You also obtain earnings from each ED allocated to the group account by others. You earn 0.4*1 = 0.4 EDs for each ED allocated to the group account by another member.

EXAMPLE 2

Relative to Example 1 assume that you decrease your allocation to 0 EDs but nothing else changes:

	Endowment	Allocation to Group Account
You	30 EDs	0 EDs
Member A	30 EDs	30 EDs
Member B	10 EDs	10 EDs
Member C	10 EDs	0 EDs

The total number of EDs in the group account = 0+30+10+0 = 40 EDs, so each group member earns = 0.4*40 = 16 EDs from the group account.

What are your period earnings in this example?
You have a 30 ED endowment and allocated 0 EDs:

> Your period earnings = (your endowment – your allocation) + (0.4*total group allocation)
> = (30 – 0) + 0.4*40
> = 30 + 16 = 46 EDs (An increase of 9 EDs relative to Example 1)

What are the period earnings of Member A in this example?
Group member A has a 30 ED endowment and allocated 10 EDs:

> Their period earnings = (their endowment – their allocation) + (0.4*total group allocation)
> = (30 – 30) + 0.4*40
> = 0 + 16 = 16 EDs (A decrease of 6 EDs relative to Example 1)

What are the period earnings of Member B in this example?
Group member B has a 10 ED endowment and allocated 10 EDs:

> Their period earnings = (their endowment – their allocation) + (0.4*total group allocation)
> = (10 – 10) + 0.4*40
> = 0 + 16 = 8 EDs (A decrease of 6 EDs relative to Example 1)

What are the period earnings of Member C in this example?
Group member C has a 10 ED endowment and allocated 0 EDs:

> Their period earnings = (their endowment – their allocation) + (0.4*total group allocation)
> = (10 – 0) + 0.4*40
> = 10 + 16 = 26 EDs (A decrease of 6 EDs relative to Example 1)

Compared with the earnings of Example 1, your earnings have increased, and the earnings of **each** of the other three members have decreased.

Comprehension 1

Please answer the following questions. Raise your hand if you need any help. A member of the experiment team will check your answers when you are done. We will begin when everyone has finished. Thank you for your patience.

1) Suppose that each group member, **including** you, allocates their *entire* endowment to the group account.

Suppose you have a 10 ED endowment and you allocate 10 EDs:

a What are your private account earnings? ________________

b What is the total number of EDs in the group account? ________________

c What are your group account earnings? ________________

d What are your period earnings? ________________

Now suppose you have a 30 ED endowment and you allocate 30 EDs:

a What are your private account earnings? ________________

b What is the total number of EDs in the group account? ________________

c What are your group account earnings? ________________

d What are your period earnings? ________________

2) Suppose that each group member, **including** you, allocates 0 EDs to the group account.

Suppose you have a **30 ED** endowment:

a What are your private account earnings? ________________

b What is the total number of EDs in the group account? ________________

c What are your group account earnings? ________________

d What are your period earnings? ________________

3) Suppose that each group member, **excluding** you, allocates 10 ED to the group account.

Suppose you have a 30 ED endowment and you allocate 0 EDs:

a What are your private account earnings? ________________

b What is the total number of EDs in the group account? ________________

c What are your group account earnings? ________________

d What are your period earnings? ________________

Assume you have a 30 ED endowment and you allocate 10 EDs:

a What are your private account earnings? ________________

b What is the total number of EDs in the group account? ________________

c What are your group account earnings? ________________

d What are your period earnings? ________________

Stage 2

In each period, after Stage 1, your earnings are initially computed will be referred to as your ***Initial Period Earnings***. You will be shown:

- Your group account allocation
- The sum of the group account allocations by all members of your group
- Your group account earnings
- Your period earnings

In Stage 2, there will be a **deductions mechanism** which may affect your period earnings.

How does the deductions mechanism affect period earnings?

In each period, after each group member has made their allocation decision, each of you will continue to be shown the individual allocations and endowments of each group member by random ID.

Each group member will now have the opportunity to assign Reduction Points to other group members. The number of Reduction Points assigned can be any integer between 0 and 10 and can be distributed in any way among group members. Note that you don't need to assign any Reduction Points and you can only assign up to 10 Reduction Points. For each Reduction Point you assign to another group member you will pay 1 ED. This cost is referred to as:

Your Administrative Costs = The number of Reduction Points *you assign to others*

For each reduction point that is assigned to you your initial period earnings will be reduced by 3 EDs. This cost is referred to as:

Your Reduction Costs = 3 * The number of Reduction Points *assigned to you from others*

To calculate your period earnings you subtract your administrative costs and your reduction costs from your initial period earnings.

Note that your period earnings cannot be negative unless you assign Reduction Points. That is, you pay Administrative Costs.

Period Earnings = Max[Initial Period Earnings – Reduction Costs, 0] - Administrative Costs

Once each member has made their decisions concerning Reduction Points you will be shown:

- Your Administrative Costs
- Your Reduction Costs
- Your Period Earnings

1

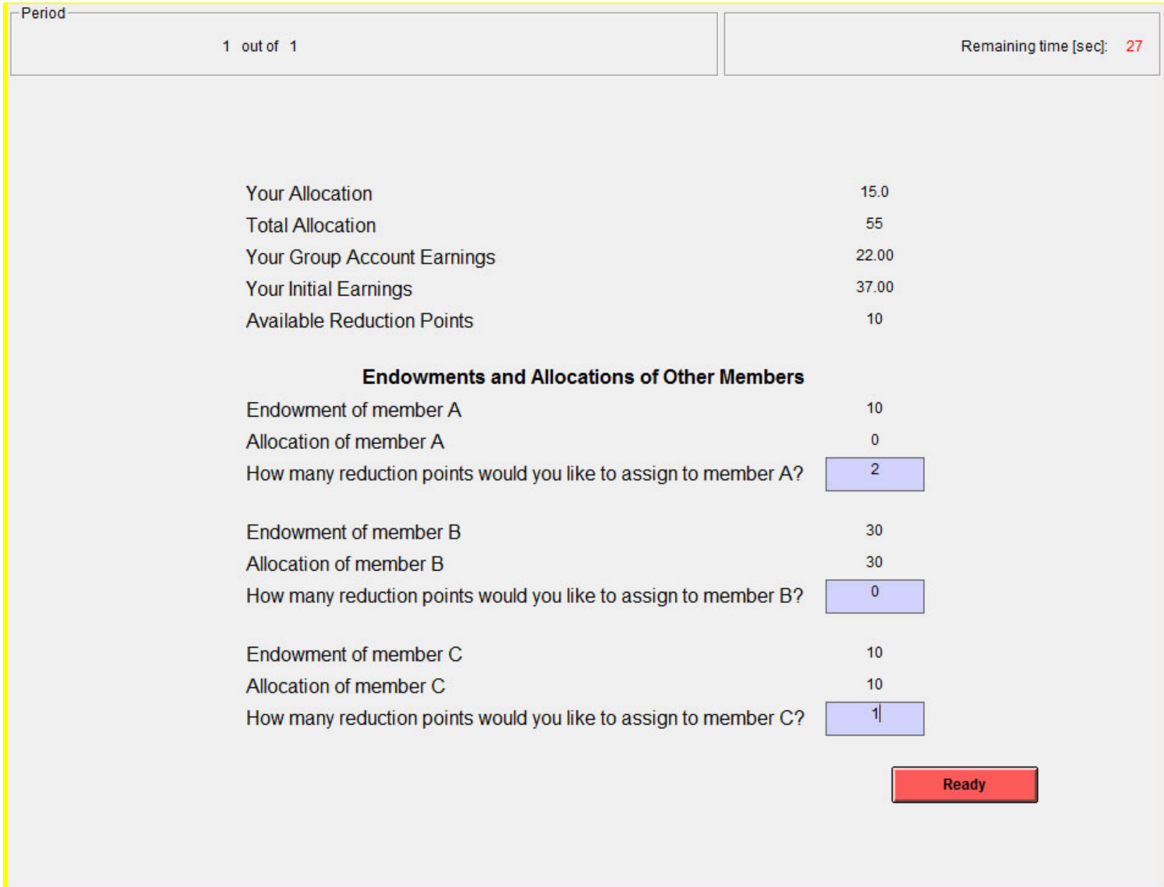


Figure 2: Example of your Reduction Point input screen given the example above.

EXAMPLE 3

The example assumes the following:

	Endowment	Allocation	Reduction Points Assigned	Reduction Points Received
You	30 EDs	15 EDs	2 to Member A 1 to Member C	1 from Member B
Member A	10 EDs	0 EDs	None	2 from you
Member B	30 EDs	30 EDs	1 to you	2 from Member C
Member C	10 EDs	10 EDs	2 to Member B	1 from you

Note that this information is provided for illustration only. You will not know how the other group members assigned their reduction points or which group members assigned reduction points to you (if any). In addition, you will not observe the endowments of other subjects, and they will not observe your endowment

The total number of EDs in the group account is = 15+0+30+10 = 55 EDs, so each group member earns = 0.4*55 = 22 EDs from the group account.

What are your period earnings in this example?
You have a 30 ED endowment, allocated 15 EDs, assigned 3 Reduction Points, and received 1 Reduction Points:

> Your initial period earnings = private account earnings + group account earnings
> = (your endowment – your allocation) + (0.4*total allocation)
> = (30 – 15) + 0.4*55
> = 15 + 22 = 37 EDs
>
> Your administrative costs = 1 ED per Reduction Point you assigned (you assigned 3)
> = 1*3 = 3 EDs
>
> Your reduction costs = 3 EDs per Reduction Point assigned to you (you received 1)
> = 1*3 = 3 EDs
>
> Your period earnings = your initial period earnings – your administrative costs – reduction costs
> = 37 – 3 – 3 = 31 EDs

1

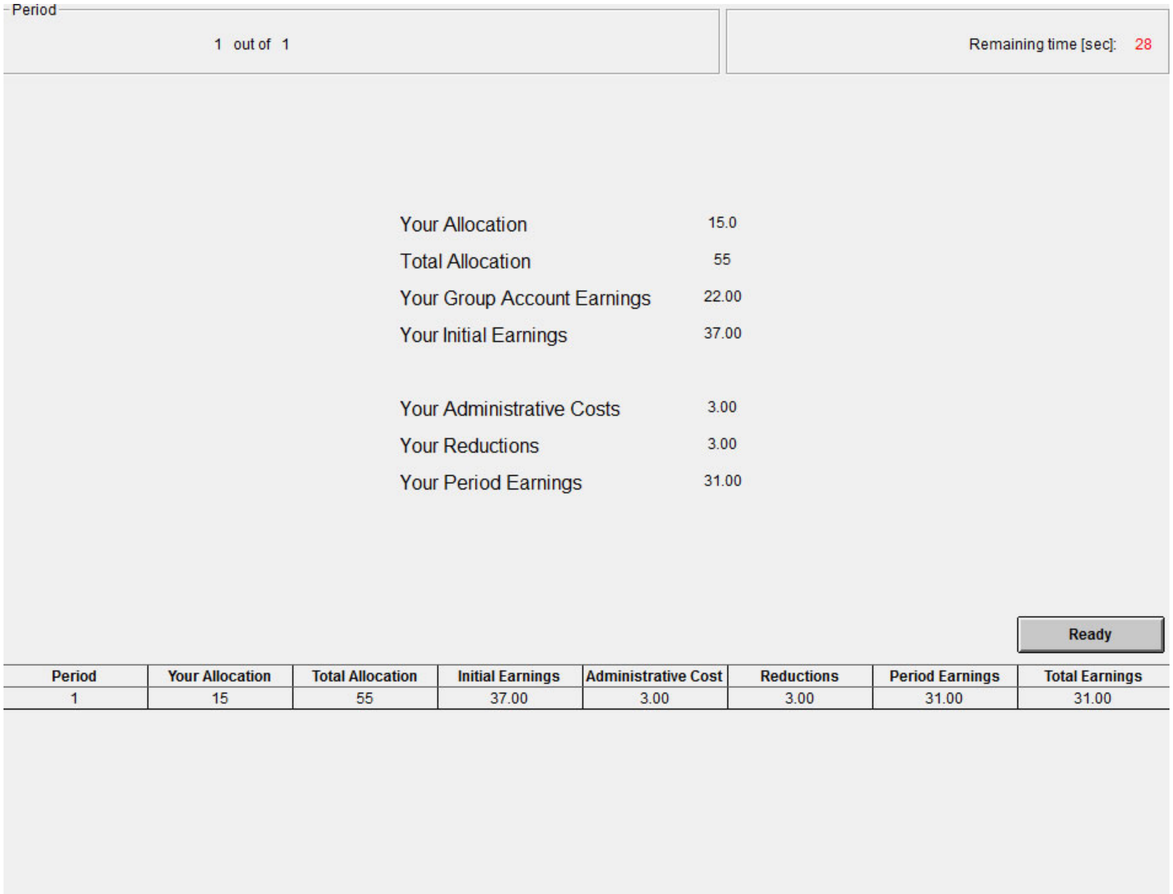


Figure 3: Example of your earnings screen given the example above

What are the period earnings of group member A in this example?
Member A has a 10 ED endowment, allocated 0 EDs, assigned 0 Reduction Points, and received 2 Reduction Points:

Member A's initial period earnings = (10 – 0) + 0.4*55
= 10 + 22 = 32 EDs

Member A's administrative costs = 1 ED per Reduction Point assigned (they assigned 0)
= 1*0 = 0 EDs

Member A's reduction costs = 3 EDs per Reduction Point received (they received 2)
= 2*3 = 6 EDs

Member A's period earnings = initial period earnings – administrative costs – reduction costs
= 32 – 0 – 6 = 26 EDs

Comprehension 2

Using the example above please answer the following questions. Raise your hand if you need any help. A member of the experiment team will check your answers when you are done. We will begin when everyone has finished. Thank you for your patience.

1. Determine the period earnings for Member B in the example above. Member B has a 30 ED endowment, allocated 30 EDs, assigned 0 Reduction Points, and received 3 Reduction Points.

   a What are Member B's initial period earning? ____________________

   b What are Member B's administrative costs? ____________________

   c What are Member B's reduction costs? ____________________

   d What are Member B's period earnings? ____________________

2. Determine the period earnings for Member C in the example above. Member C has a 10 ED endowment, allocated 0 EDs, assigned 2 Reduction Point, and received 2 Reduction Points.

   a What are Member C's initial period earning? ____________________

   b What are Member C's administrative costs? ____________________

   c What are Member C's reduction costs? ____________________

   d What are Member C's period earnings? ____________________