Punishment with Imperfect Monitoring reduces Cooperation and Amplifies Inequality in a Dynamic Public Goods Game

Abstract

Cooperation in social dilemmas like the public goods game (PGG) is often sustained by punishment, yet real-world enforcement rarely operates under perfect information about others' contributions. We use a dynamic PGG framework to examine how imperfect monitoring of contributions and costly peer punishment interact to shape cooperation, wealth accumulation, and inequality. In our 10-round design, participants' earnings carry over from one round to the next, capturing the compounding nature of many real-world public goods problems. We introduce two key features: a peer punishment mechanism where participants can spend 1 token to reduce another group member's earnings by 3 tokens, and imperfect monitoring, where contribution decisions are underreported with a probability of 25%. We find that punishment alone increases average contributions, but it does not improve overall wealth and, in fact, amplifies inequality. When punishment is combined with imperfect monitoring, it leads to substantial inefficiencies – cooperation falters, wealth shrinks, and inequality grows. These effects are driven in part by the rise of antisocial punishment, i.e., free-riders targeting cooperators, undermining incentives to contribute. Our results underscore a critical trade-off: punishment can support cooperation when contributions are perfectly observable, but under noisy conditions, it may generate unintended consequences that erode welfare.

Keywords Dynamic public goods game, peer punishments, imperfect monitoring, inequality
Classification Codes C92, D63, H41

1. Introduction

The widespread ability of humans to cooperate with strangers, even when self-interest offers clear material advantages, presents a puzzling paradox. This has sparked interest from researchers across various disciplines like political science (Ostrom, 1998; Ostrom et al., 1992), evolutionary biology (Axelrod, 1984; Gintis, 2000; Milinski et al., 2002; Nowak, 2006; West et al., 2007), anthropology (Hill et al., 1993; Stibbard-Hawkes et al., 2022), sociology (Coleman, 1988; Marwell & Ames, 1979), psychology (Henrich & Muthukrishna, 2021; Rand & Nowak, 2013) and economics (Acemoglu & Jackson, 2015; Ostrom, 2015). Economists frequently use the public goods game (PGG) paradigm to study cooperation and its sustainability. In the canonical linear PGG (Ledyard, 1995), participants receive a fixed endowment and choose how much of this endowment to contribute to a communal pot, which is then multiplied and redistributed equally among players. While full cooperation maximizes collective payoff, free-riding, which emerges as the dominant strategy, undermines cooperation. This tension mirrors real-world challenges, such as funding public infrastructure or addressing climate change, where collective gains depend on individuals resisting individual incentives to free-ride.

To better understand the persistence of cooperation, researchers have extended the PGG to repeated interactions, where participants play the same game over multiple rounds. In the standard repeated PGG, participants start each round with the same endowment and face the same payoff structure, effectively resetting the game to a fresh state. This structure limits their ability to capture the intertemporal dependencies characteristic of many real-world public goods. In response, researchers developed dynamic PGGs where payoffs in a given round are endogenously linked to outcomes from previous rounds. One prominent approach involves allowing participants to carry over their accumulated earnings as endowments for subsequent rounds (Gächter et al., 2017; Rockenbach & Wolff, 2019). This structure captures the potential for reinvestment and compounding returns, offering a richer framework for studying the evolution of cooperation, wealth accumulation, and inequality. Although free-riding remains the dominant strategy, dynamic PGGs reveal the potential for compounding benefits from sustained cooperation. We use “dynamic PGG” to refer to this class of games with evolving state variables, in contrast to static PGGs where key parameters remain fixed.

Studies using static PGGs have shown that costly peer punishment can help sustain cooperation (Fehr & Gächter, 2000, 2002; Sefton et al., 2007). Although such punishment reduces net group payoffs in the short run due to the punishment cost, evidence from longer time horizons suggests that it can

ultimately improve overall group welfare by stabilizing cooperative norms and deterring defection (Gächter et al., 2008). Equally, other studies have highlighted that the effectiveness of punishment in static PGGs in terms of increasing cooperation crucially depends on the ability to perfectly monitor others' contributions (Ambrus & Greiner, 2012, 2019; Grechenig et al., 2010). Most experiments assume perfect information, allowing individuals to target defectors accurately. Yet when feedback is noisy or incomplete, as it often is in real-world settings, punishment loses its deterrent power. Misidentification leads to misdirected sanctions, undermining trust and cooperation rather than reinforcing them.

In this paper, we examine how the combination of punishment and imperfect public monitoring affects contributions, wealth generation, and inequality in a dynamic setting. We design a 10-round PGG where participants' earnings accumulate over time, with each round's earnings serving as the starting endowment for the next. This design mimics the compounding effects of individual decisions on wealth[1] and allows us to observe the evolution of inequality within groups. To this framework, we add two key features. First, a costly peer punishment mechanism allowing participants to penalize others by deducting 3 tokens at a personal cost of 1 token. Second, we introduce noise in the monitoring of contributions: with a 25% chance, contributions to the public good are underreported.[2] Our design includes four treatments: a baseline with neither punishment nor monitoring noise, and three variants – punishment only, noise only, and a combined treatment featuring both. To our knowledge, this is the first study to examine how these mechanisms interact in a dynamic wealth-building context.

Our results replicate key findings from Gächter et al. (2017) and Rockenbach & Wolff (2019), confirming that peer punishment in dynamic public goods settings increases average contributions, both in absolute tokens and as a share of endowments. However, this headline effect conceals substantial heterogeneity. A small number of cooperative groups drive the aggregate increase, while most groups exhibit stagnation or even decline. Importantly, higher contributions under punishment do not translate into higher wealth accumulation. The efficiency gains from increased cooperation are more than offset

[1] For the context of this study, we have defined "wealth" similarly to Gächter et al. (2017). The wealth sums the endowment of all participants in a given group at the beginning of the following period. In other words, wealth in round t will be the amount of tokens generated in round $t-1$.

[2] Following the approach of Ambrus & Greiner (2012), we incorporate noise in the information about contributions in only one direction – underreporting, not overreporting. Their rationale is that in real-life partnerships, it is more common for members to contribute to a joint project without others recognizing their contributions, rather than non-contribution being mistaken for contribution. However, unlike their study, we increased the probability of noise from 10% to 25%, to make the effect more salient.

by the costs of sanctioning, resulting in significantly lower final-round wealth in punishment treatments relative to the baseline.

Beyond this, our findings reveal how imperfect monitoring, alone and in combination with punishment, shapes outcomes in dynamic settings. When implemented in isolation, imperfect monitoring exerts no meaningful effect on contributions, earnings, or inequality. But when paired with punishment, it generates sharp inefficiencies: contributions fall, total wealth shrinks, and inequality rises. This deterioration stems in part from a surge in anti-social punishment (penalizing cooperators) likely triggered by misperceptions arising from noisy feedback. Analysis of within-group Gini coefficients reveals that both punishment treatments, whether implemented alone or with imperfect monitoring, generate significantly higher inequality relative to the baseline and the noise-only condition.

Together, these findings extend the existing literature by identifying key boundary conditions for the effectiveness of peer punishment in dynamic settings. Specifically, we show that peer punishment fails to enhance either efficiency or equity when information is imperfect and group behaviour is heterogeneous. This underscores the need to account for informational frictions and endogenous wealth dynamics when evaluating the institutional robustness of decentralized enforcement mechanisms. The remainder of the paper is structured as follows: Section 2 presents a selective literature review; Section 3 describes the experimental setup and procedures; Section 4 introduces the hypotheses; Section 5 presents the results; and Section 6 offers concluding remarks.

2. Literature Review

2.1. Dynamic Public Goods Games: A Comprehensive Review

To better understand how cooperation can be sustained over time, researchers extended the one-shot PGG into repeated-play settings (see Chaudhuri, 2011 and Ledyard, 1995 for review). In standard repeated PGGs, each round begins with the same endowment and payoff structure. While this format allows for learning, reciprocity, and strategic punishment, it fails to capture a central feature of real-world public goods: intertemporal dependency. In many real-world settings, today's contributions directly influence the resources and incentives available tomorrow.

To address this gap, researchers have explored a dynamic variation of the repeated PGG, where outcomes of each round influence the payoff structure in subsequent rounds. In this dynamic PGG, participants typically engage in multiple rounds with fixed group members. However, unlike the static

version, the dynamic game no longer resets in each round. Instead, payoff-relevant parameters (such as the endowment or the MPCR) evolve endogenously based on past outcomes. This inter-round dependency creates a cascading structure in which today's decisions influence tomorrow's opportunities, reflecting the endogenous dynamics of many real-world collective action problems. Consider, for example, a small town investing in community infrastructure. Early contributions might fund the construction of a community centre, which not only improves local amenities but also generates revenue that can be reinvested in further community projects. Similarly, on a broader scale, early investments in renewable energy can drive down future costs and foster additional technological breakthroughs. In both examples, initial spurts of cooperation can create a self-reinforcing cycle that generates compounding welfare. These dynamic settings better reflect reality, where today's investments shape tomorrow's outcomes.

The literature highlights two main mechanisms through which these temporal linkages operate. First, *endogenous return rate* allows the MPCR to evolve based on the group's past contributions (Cadigan et al., 2011; Noussair & Soo, 2008; Sadrieh & Verbon, 2006). The second, and the focus of our study, ties current payoffs to future endowments through *wealth accumulation*. In this framework, a player's earnings at the end of one round determine their starting endowment in the next, much like an investment portfolio, where returns from one period add on to the principal for the next. This approach not only tracks what happens to contributions but also to the accumulation of wealth and within-group inequality.

Gächter et al. (2017) used the wealth accumulation approach to examine whether cooperation can be sustained in a dynamic setting, how endogenous inequality influences long-term wealth distribution, and whether punishment encourages or stifles overall contributions.[3] In their design, participants were assigned to fixed groups of four and they played for either 10 or 15 rounds. As their design had wealth accumulation, it created compounding effects: groups that started with higher cooperation experienced sustained wealth accumulation, while those with early free-riding tendencies saw their resources dwindle. A key feature of their study was the introduction of peer punishment: participants could reduce another participants' wealth by 3 tokens at a cost of one token to themselves. Theoretically, punishments could enforce cooperation by deterring free-riding.

[3] Our study is most directly inspired by Gächter et al. (2017), whose experimental design shares key similarities with ours.

Their findings revealed three important results. First, even in the absence of punishment, absolute contributions increased over time, a stark contrast to static public goods experiments where contributions typically decline. However, when measured as a proportion of endowment, contributions declined over time, similar to static games. This suggests that the growth in absolute contributions was primarily driven by increasing endowments, not stronger intrinsic cooperative norms. Second, there was substantial variation in wealth accumulation across groups. The richest groups amassed more than ten times the wealth of the poorest, highlighting the self-reinforcing nature of early cooperative behaviour. Third, within-group inequalities emerging in the initial rounds had long-term adverse consequences. Groups that developed high inequality in the first few rounds often remained trapped in a low-growth equilibrium, particularly when punishment was an option. In these cases, punishment power was unequally distributed, with wealthier participants potentially able to punish more effectively than poorer ones. The study found that in high-inequality groups, antisocial punishment (where low contributors punished high contributors) was more prevalent, and this pattern often reinforced disparities rather than promoting cooperation.

In a more recent study, Rockenbach & Wolff (2019) examined how punishment influences cooperation in a dynamic PGG with accumulated earnings. Their central question was whether punishment would reinforce cooperation over time or instead provoke destructive retaliation cycles that undermine group welfare. Conducted over 20 rounds with fixed four-person groups, their experiment featured two treatments: one followed a standard dynamic PGG with wealth accumulation, similar to Gächter et al. (2017), while the other added a second stage allowing for peer punishment. The punishment mechanism was convex, reflecting real-world power asymmetries, higher penalties became increasingly expensive.[4] Their findings indicate that while punishment can increase contributions relative to endowments, it does not reliably translate into higher net earnings. In the punishment treatment, participants contributed a greater share of their available wealth, indicating that the mere availability of sanctions can promote prosocial behaviour. However, the expenses of punishing and being punished frequently offset the gains from higher contributions. As a result, average final wealth levels were not significantly higher in the punishment treatment.

[4] The cost of punishment was calculated as one-third of the punishment amount plus the cube of that amount divided by 2000, making larger punishments disproportionately more expensive.

A key mechanism driving these outcomes was the pattern of punishment use in early rounds. Early, aggressive punishment often provoked retaliation, triggering costly cycles of mutual sanctioning that depleted group resources. Only groups that applied punishment cautiously and avoided retaliation were able to maintain high cooperation and accumulate substantial wealth. Moreover, while initial contribution levels strongly predicted group success in the no-punishment treatment, consistent with findings from static games (e.g., Fischbacher et al., 2001), they were not predictive in the punishment condition. Instead, early punishment dynamics, particularly the presence or absence of retaliatory behaviour, were the key determinants of long-term outcomes.

Taken together, these findings underscore one important differences between static and dynamic PGG settings. In static games, the availability of peer punishment significantly increases contributions to the public good (Andreoni & Gee, 2012; Fehr & Gächter, 2000, 2002). However, the impact on net earnings remains ambiguous: while punishment promotes cooperation, it also incurs direct costs for both punishers and recipients. Gächter et al. (2008) showed that the net payoff from punishment is time-sensitive: negative over short horizons (10 rounds) but positive over long horizons (50 rounds) as cooperation becomes entrenched. Yet in dynamic settings with endogenous wealth, these theoretical benefits are undermined. In Gächter et al. (2017), peer punishment failed to increase contributions relative to endowments and significantly reduced accumulated wealth in the 15-round condition. Similarly, Rockenbach & Wolff (2019) showed that although punishment increased relative contributions, it did not raise final earnings on average. Instead, it often triggered retaliatory sanctioning cycles that eroded group wealth.

Contribution dynamics also differ markedly across the two settings. In static PGGs, contributions decline over time, typically falling from roughly half of the initial endowment to near 10 per cent (Chaudhuri, 2011; Ledyard, 1995). Dynamic games yield more heterogeneous outcomes. Some groups maintain high levels of cooperation and accumulate substantial wealth, while others descend into persistent low-contribution equilibria. These divergent trajectories are driven in part by the compounding nature of dynamic incentives: early cooperation can generate sustained wealth growth, whereas early free-riding may trap groups in long-term under provision.

These findings underscore the importance of designing public goods systems that account for intertemporal incentives. Since dynamic PGGs better capture real-world decision-making, more research is needed to understand how different interventions shape cooperation, wealth, and inequality.

One factor that may be crucial in the context of dynamic PGGs is the accuracy of the information available. Such accuracy or lack thereof may have significant impact on cooperation in dynamic PGGs. We examine the evidence here next.

2.2. Information Accuracy and Punishment in PGGs

A large body of experimental research using static PGGs, beginning with Fehr and Gächter's (2000, 2002) seminal work, shows that costly punishment can sustain high contribution levels when contributions are perfectly observable. In noise-free environments, defectors are immediately identified and can be sanctioned, which deters free-riding, at least within the static framework. Most laboratory studies assume this setting of error-free observation. Yet in the real-world, monitoring is rarely so precise; individuals often have only partial information about others' contributions. What happens when monitoring is imperfect?

Detection accuracy is expected to critically impact deterrence: if potential free-riders know there's a good chance their shirking will go unnoticed, punishment loses its bite. And if cooperators fear being punished in error, cooperation may completely break down. For instance, in a team project where individual effort is hard to verify, a member might shirk, hoping others won't notice, while diligent teammates risk being wrongly blamed. Tax compliance offers a parallel: since audits cover only a fraction of returns, low detection risk can invite evasion. Thus, enforcement faces two risks: false positives (type 1 error), where cooperators are penalized, and false negatives (type 2 error), where free-riding escapes sanction. Both undermine cooperation: undetected free-riders exploit the group, while mistaken punishment discourages effort. In this way, imperfect monitoring not only weakens the deterrent power of punishment, it destabilizes the entire system of mutual accountability on which cooperation rests. Below, we survey key studies that explore this phenomenon.

Grechenig et al. (2010) showed that the use of punishments become significantly less effective at sustaining cooperation when monitoring was imperfect. In a 10-round static PGG with four players holding 20-token endowments each, participants could use costly peer punishment, deducting 3 points from another's payoff at a cost of 1. The key treatment varied monitoring accuracy. Under perfect monitoring, contributions were observed exactly. In contrast, under imperfect monitoring, signals reflected the true value of contributions with probability λ, and with probability $1-\lambda$, the observed contribution was replaced by a randomly drawn integer between 0 and 20, excluding the true value. These random errors were applied independently across all observers and targets. This meant that a

player could be misjudged by one peer but correctly assessed by another, or misjudged by all. The authors tested two noise levels: a low-noise condition ($\lambda = 0.9$), and a high-noise condition ($\lambda = 0.5$). In both conditions, punishment decisions were based solely on potentially inaccurate signals.

Under perfect monitoring, punishment proved highly effective: average contributions nearly doubled, rising from 9.2 tokens without punishment to 16.8 tokens with it. However, once noise is introduced, the dynamics changed. Under the high-noise condition, the average contribution with punishment fell to 11.3 tokens, statistically indistinguishable from the 8.8 tokens observed when punishment was unavailable. The low-noise condition did not statistically differ from the perfect monitoring baseline. Thus, while punishment sustained cooperation under perfect or near-perfect monitoring, its effectiveness deteriorated sharply once signal accuracy dropped to 50%, where identifying free-riders becomes unreliable. The results further showed that under noisy conditions, participants remained eager to punish, often punishing the wrong targets. These type 1 errors, punishing cooperators due to misleading signals, carried a double cost: they imposed losses on innocent players while failing to correct behaviour. Type 2 errors also emerged, as true free-riders sometimes avoided punishment due to inaccurate signals. The coexistence of both error types meant that sanctions were not only misallocated but systematically misaligned with actual behaviour.

Ambrus & Greiner (2012) conducted a 50-round static public goods experiment comparing perfect versus imperfect public monitoring in combination with different punishment severities. Each group comprised three participants who repeatedly made binary contribution decisions, either contributing their entire 20-token endowment to a group account or withholding it. In the noisy setting, a true contribution was publicly misreported as defection with a 10% probability, while actual defections were always recorded accurately. This asymmetry in noise, where only cooperators faced a risk of being misclassified as free-riders (type 1 error), differentiates this design from Grechenig et al. (2010), where errors could occur in both directions. Additionally, Ambrus & Greiner (2012) implemented a *public* monitoring system in which all participants observed the same, potentially erroneous, signal – another difference from Grechenig et al. (2010), whose participants observed individually varying signals. Players were fully informed of this error structure. To test the interaction of noise with punishment severity, the authors implemented three punishment regimes: no punishment, standard punishment (3:1 damage-to-cost ratio with capped harm to prevent negative payoffs), and strong punishment (6:1 ratio with no cap).

Under perfect monitoring, the results aligned with prior literature: the option to punish led to near-complete cooperation, and reaped significant net earnings gains. With standard punishment under imperfect monitoring, however, sanctions were frequently applied but largely ineffective. Groups failed to eliminate free-riding, and net earnings fell below even the no-punishment baseline, as the costs of punishment outweighed any cooperation gains. In effect, punishment under noisy monitoring imposed costs without delivering the cooperation dividend. Introducing strong punishment partially mitigated this inefficiency. Contributions rose substantially, and the threat of more severe sanctions helped sustain cooperation despite noise. However, the welfare gains from increased cooperation merely offset the higher cost of punishing, resulting in net earnings statistically indistinguishable from the no-punishment baseline.

To sum, the study found a U-shaped relationship between punishment severity and group payoffs under noise. Standard punishment led to the lowest earnings, while both no punishment and strong punishment produced similar, relatively higher payoffs. The authors traced this to two key responses. First, contributors who were mistakenly punished (type 1 error) became less likely to contribute in future rounds. Second, the deterrent effect of punishment on actual defectors was weakened by uncertainty about whether sanctions were truly deserved. The strong punishment partially restores the deterrence effect of punishment, but only at a higher cost. Taken together, both Grechenig et al. (2010) and Ambrus & Greiner (2012) converge on a policy implication: before relying on punishment, institutions must first ensure high-fidelity information; otherwise, even well-intentioned sanctioning mechanisms can erode rather than enhance collective welfare.

2.3. Key Insights and Research Gaps

The experimental literature on PGGs with punishment under imperfect monitoring converges on a clear insight: punishment is less effective in curbing free-riding without reliable information. Two patterns emerge. First, noise introduces both type 1 and type 2 errors, but it is the risk of punishing cooperators (type 1 error), that most strongly undermines efficiency (Markussen et al., 2016). Unjustly punished contributors tend to reduce their future contributions, while potential punishers hesitate under uncertainty, weakening deterrence. Second, institutional design matters. While all punishment regimes suffer under noisy conditions, mechanisms involving collective decisions, such as democratic punishment (Ambrus & Greiner, 2019) or threshold-triggered group sanctions (Mengel et al., 2021), can partially offset the damage by curbing arbitrary retaliation and rooting enforcement in shared norms.

A gap in this literature is its predominant focus on static PGGs, where decisions and outcomes are confined within each round. This suggests scope for future research exploring the interaction between imperfect monitoring and punishment in dynamic PGGs, where payoffs accumulate over time. This coupling of dynamic incentives and imperfect monitoring introduces new complexities. In contrast to static games, where punishment errors are self-contained within each round, dynamic environments are path-dependent: early misclassifications can have lasting consequences by altering future endowments or wealth. Since wealth determines punishment capacity, early errors can entrench asymmetries, giving more power to the wealthy, leading to an inequality not just in terms of endowment but also in terms of punishing power. Furthermore, while static games typically involve immediate reactions to noisy signals, dynamic environments allow players to develop forward-looking strategies, weighing the long-term costs of punishing. Players may face a second-order dilemma: although punishing a suspected free rider may seem justified in the moment, doing so reduces that player's future contribution potential, which could harm welfare in the long run. This compounds the opportunity cost of misdirected punishment. Over time, participants may also become more sophisticated in processing noisy information, using past behaviour and patterns of misclassification to adjust their beliefs about others, which could potentially increase tolerance towards free-riders.

3. Experimental Setup

3.1. Experimental Procedure and Treatments

At the start of the experiment, participants were randomly assigned to groups of three, with group composition fixed throughout all 10 rounds. Participants knew they were interacting with the same group members in each round, but the identities of these members ("Other 1" and "Other 2") were kept confidential. We also informed them at the outset that their decisions would span 10 rounds. Each participant was allotted 'tokens' to use in these rounds, and they were aware that these tokens would be converted to real money at the end of the experiment, with an exchange rate of 25 tokens equalling £1. We then divided the groups into four treatments, based on a 2 x 2 factorial design that varied by two factors: the punishment mechanism and the monitoring mechanism. This leads to the following four treatments.

3.1.1. Baseline Condition

Each participant received 20 tokens as an initial endowment. In round 1, participants decided how many tokens to contribute to a 'group account', with the remainder automatically allocated to their 'private account'. The group account contributions were pooled, multiplied by 1.5, and then equally divided among the three members. Tokens kept in the private account remained unchanged and were added to each participant's earnings for that round. After making their decisions, participants viewed a 'public record' displaying the contributions of all group members. The screen also showed the earnings of each participant at the end of the round. The same process was repeated in subsequent rounds, with the tokens earned at the end of one round becoming the endowment for the next. Formally, for a participant i with an endowment E_i^t who contributed C_i^t tokens to the group account, their total earnings at the end of the round t (which became their endowment for round $\mathrm{t}+1$) is:

$$E_i^{t+1} = E_i^t - C_i^t + \frac{1.5}{3}\sum_{j=1}^{3} C_j^t$$

To ensure understanding, we provided detailed instructions and examples of the game, followed by three comprehension check questions.[5]

3.1.2. Peer Punishment Treatment

The punishment treatment maintained the structure of the baseline condition but with costly peer punishment. After viewing the public record and their earnings, participants could penalize other group members at a personal cost. Specifically, they could deduct 3 tokens from any group member at a cost of 1 token to themselves. This design allowed for targeted punishment of specific individuals. To prevent negative earnings, we did not execute punishments that would reduce a participant's earnings below zero. Formally, we calculated the earnings for participant i at the end of round t and hence, at the start of round $\mathrm{t}+1$ as:

$$E_i^{t+1}max = \left\{ E_i^t - C_i^t + \frac{1.5}{3}\sum_{j=1}^{3} E_j^t - \sum_{j=1\neq i}^{2} P_{ji}^t - \frac{1}{3}\sum_{j=1\neq i}^{2} P_{ij}^t,\ 0 \right\}$$

[5] In the comprehension check questions, we provide hints of how the group account returns are calculated and provide them with an on-screen calculator to help with the calculation.

Where $\sum_{j=1\neq i}^{2} P_{ji}^{t}$ represents the punishment tokens deducted by other group-members from participant i and $\frac{1}{3}\sum_{j=1\neq i}^{2} P_{ij}^{t}$ represents the cost of punishment for participant i. In addition to the baseline comprehension checks, we included two questions about the punishment mechanism.

3.1.3. Imperfect Monitoring Treatment

This treatment adds noise to the baseline condition. While the public record displayed how many tokens each member contributed to the group account, there was a 25% chance that a participant's contribution would be shown as lower than the actual amount. Specifically, for a participant i with an endowment E_i^t who contributed C_i^t tokens to the group account, the public record will display:

$$Public\ record\ = \begin{cases} C_i^t & with\ probability\ 0.75 \\ U[0, C_i^t - 1] & with\ probability\ 0.25 \end{cases}$$

For example, if a participant contributed 10 tokens, with a 75% chance it would display as 10, and a 25% chance it would display any whole number from 0 to 9 with equal probability. These errors did not affect actual payoffs, and thus, the wealth equation remains structurally identical. However, the noise in the monitoring system introduced uncertainty in the perception of participants that could influence decisions in the subsequent rounds. The remaining procedure mirrored the baseline condition, with end-of-round earnings becoming the next round's endowment. To ensure understanding, we included an additional comprehension question about the imperfect monitoring mechanism.

3.1.4. Combined Treatment (Peer Punishment plus Imperfect Monitoring)

In this treatment, participants made their contribution decisions as in the baseline condition, but the public record displayed contributions with the same possibility of error as in the imperfect monitoring treatment. After viewing this record, participants could choose to penalize other group members, just as in the punishment treatment.[6] We included comprehension check questions from both the punishment and imperfect monitoring treatments to ensure participants fully understood the combined structure.

[6] Note that in this treatment, participants can infer others' contributions, but doing so is cognitively demanding under time constraints. Accurately identifying individual contributions (especially when conditioning future punishments) requires not only quick calculation but confidence in one's assessment when it diverges from the public record. This is possible but not easy and therefore, we believe that this treatment provides a good proxy for imperfect monitoring. Unlike Ambrus & Greiner (2012), we must display individual earnings because accumulated earnings determine future endowments. To ensure consistency, participants receive accurate information about their own earnings and endowments at the start of each round, even if the public record is noisy.

As noted earlier, the dynamic nature of the PGG allows for exponential growth in collective wealth. If each participant, starting with an endowment of 20 tokens, contributes their entire endowment to the public good in every round and avoids punishing others, a group of three can accumulate a total of 2,306.6 tokens by the end of 10 rounds. This total equates to approximately £92.26, demonstrating the substantial collective earnings achievable through maximized cooperation.

After completing the 10 rounds, participants completed a questionnaire. This included demographic questions on age, gender, nationality, education level, and employment status, as well as questions about their previous participation in economic experiments. They then answered a simple risk assessment on an 11-point Likert scale, indicating their willingness to take risks, and rated how well they understood the instructions on a similar scale. An open-ended comments section allowed for additional feedback. The questionnaire concluded with shortened versions of the Social Dominance Orientation (SDO) scale (Pratto et al., 1994) and the Right-Wing Authoritarianism (RWA) scale (Altemeyer, 1981). Higher SDO scores suggest greater tolerance for intergroup inequality, while higher RWA scores indicate a preference for social conformity. Upon completing the questionnaire, the study ended, and participants were compensated based on their final token earnings from Round 10. Figure 1 provides an overview of the experimental design.

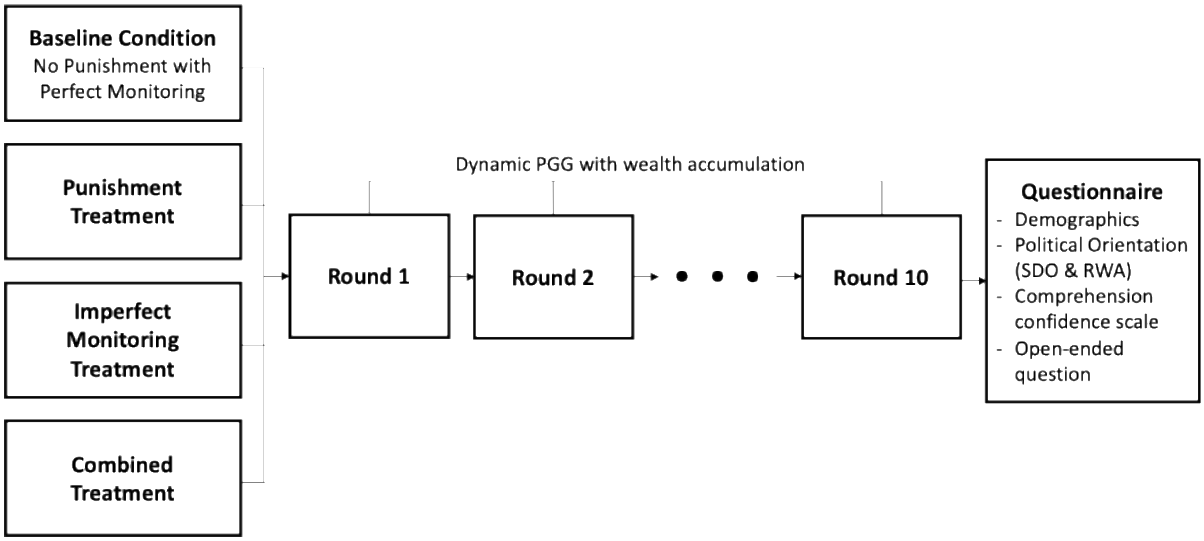

Figure 1 Overview of the experimental design

3.2. Participants

We recruited a sample of 450 participants via Prolific, a UK-based online crowdsourcing platform. These participants were organized into 150 triads. Data collection was conducted over three sessions between February and March 2024, utilizing oTree to facilitate all interactions (Chen et al., 2016). Participants

were required to be at least 18 years old, with 53% identifying as female and an average age of 29.6 years. This diverse sample allows us to make broader inferences about how people generally respond to social dilemmas when faced with a punishment mechanism and imperfect monitoring of contributions. Participants were given a time limit of 30 seconds per round for participants to decide how many tokens to allocate to the group account. In the punishment treatments, participants had an additional 45 seconds per round to decide how many tokens to deduct from each of the other two group members. If participants did not submit their choices within the time limit, a random choice was submitted on their behalf. If this occurred twice, the participant was removed from the study.[7] The University of Auckland Human Participants Ethics Committee (UAHPEC) approved the experiment (Ref. No.: UAHPEC26548). Table 1 provides summary statistics on the number of participants and their final earnings.

Table 1 Participant earnings across treatments

Treatment	Participants	Groups	Average Final Earnings	Median Final Earnings
Baseline	120	40	£4.07	£2.4
Peer Punishment	111	37	£4.43	£0.52
Imperfect Monitoring	105	35	£4.2	£2.72
Combined	114	38	£0.56	£0.32
All	**450**	**150**	**£3.42**	**£2**

4. Hypotheses

Below, we present our hypotheses, which focus on two key aspects of the dynamic PGG: contributions to the public good, wealth accumulation over the 10 rounds. Next, we pose an open-ended exploratory question concerning the effects of imperfect monitoring alone as well as its combination with punishment on within-group inequality that emerges endogenously.

4.1. Contributions to the public good

4.1.1. H1a. Imperfect monitoring without punishment will lead to lower contributions compared to the baseline scenario.

Imperfect monitoring introduces uncertainty by distorting observed contributions in one direction, actual contributions may appear lower, but never higher. This asymmetry reduces trust, as cooperative participants risk being seen as free-riders, discouraging future contributions. Free-riders, in turn, face

[7] In our analysis, we focused exclusively on the 'perfect' groups – those in which no participants timed out or dropped out.

less accountability. These mechanisms align with empirical findings by Ambrus & Greiner (2012), who demonstrated that imperfect monitoring leads to both lower contributions and a more rapid decline in cooperation in static PGGs.

4.1.2. H1b. The combination of punishment and imperfect monitoring will result in lower contributions compared to the baseline scenario.

Punishments provide a deterrent against free-riding by imposing costs on those who contribute less, thereby encouraging higher cooperation (Chaudhuri, 2011). Gächter et al. (2017) demonstrated that punishment enhances cooperation in the dynamic PGG setting, particularly in later rounds. We anticipate similar outcomes in our study with only punishment. However, when punishment is combined with imperfect monitoring, misdirected and counter punishments become more likely. Cooperative participants might reduce their contributions to avoid the risk of unfair punishment, while free-riders are less deterred when monitoring is unreliable. This creates a cycle where cooperative behaviour is discouraged and non-cooperative behaviour is incentivized. Theoretical predictions emphasise how noise amplifies recognition penalties and reduces conditional cooperation, leading to lower overall contributions, consistent with findings from Ambrus & Greiner (2012) in static PGGs.

4.2. Wealth Accumulation

4.2.1. H2a. Imperfect monitoring without punishment will lead to lower wealth generated compared to the baseline.

As reduced contributions resulting from imperfect monitoring diminish the public good's returns, the group's collective wealth declines. Lower contributions directly translate to lower wealth.

4.2.2. H2b. The combination of punishment and imperfect monitoring will result in lower wealth compared to the baseline scenario.

While peer punishment can enhance contributions, their costs, coupled with misdirected sanctions under imperfect monitoring can more than offset these gains. Gächter et al. (2017) observed negligible net effects of punishment on wealth in 10-round dynamic PGGs, with adverse outcomes in 15-round treatments. Frequent punishments consume additional resources, especially if done in the initial rounds, resulting in average wealth levels that are comparable or even lower to those observed without

punishment. When punishment is combined with imperfect monitoring, we anticipate it to significantly reduce wealth by triggering a negative feedback loop: misdirected punishment reduces contributions, which in turn escalates punishment as participants attempt to enforce cooperation. The compounded negative effects of imperfect monitoring and costly, misdirected punishments can lead to the poorest wealth outcomes in this treatment.

4.3. Inequality

4.3.1. H3a: Imperfect monitoring without punishment will lead to higher within-group inequality compared to the baseline scenario.

When monitoring is imperfect, the resulting uncertainty causes participants to vary their contributions more widely than when information is reliable. Ambrus & Greiner (2012) observed that imperfect monitoring led to more heterogeneous contribution decisions, with some individuals choosing to contribute while others did not, unlike situations with reliable information where contributions were more polarized, most participants either contributed or did not.[8] This variability in contributions increases wealth disparities, as those who contribute less benefit at the expense of those who contribute more, ultimately leading to higher inequality within the group.

5. Findings

We organize our findings in two sections. The first section provides a descriptive analysis, focusing on contributions to the public good, the trajectory of wealth accumulation over 10 rounds, and the evolution of within-group inequality. We complement this analysis with non-parametric tests and regression models. The second section explores the structure of punishment, differentiating between pro-social and anti-social punishment with punishment with and without imperfect monitoring. Here, non-parametric tests and regression analyses further clarify significant trends and treatment differences. Key statistics across treatments and all rounds are summarized in Table 2, providing an overview of the primary outcomes.

[8] In their study, Ambrus & Greiner (2012) offered participants only a binary choice: whether or not to contribute. In our study, participants can choose the exact amount of their endowment to contribute to the public good, enabling a more precise analysis of their decisions.

Table 2 Average token contributions, contribution rates (as a share of endowment), wealth, Gini-coefficients, and punishments across treatments over all rounds

Treatment	Participants / Groups	Contribution (in tokens)	Contribution Rate	Wealth (in tokens)	Gini Coefficient (in tokens)	Punishments (in tokens)
Baseline	120 / 40	16.49 (38.62)	30.56% (29.26)	59.02 (63.81)	0.10 (0.01)	-
Punishment	111 / 37	22.73 (71.05)	42.25% (35.23)	44.61 (109.17)	0.15 (0.02)	5.55 (4.29)
Imperfect Monitoring	105 / 35	15.70 (31.33)	32.80% (31.01)	58.08 (60.86)	0.13 (0.03)	-
Combined	114 / 38	7.21 (10.55)	38.17% (33.27)	21.58 (27.67)	0.19 (0.04)	7.56 (3.36)

Note: Figures in parentheses are standard deviations

5.1. Public Goods Contributions, Wealth Dynamics, and Inequality Trends

5.1.1. Contributions to the Public Good

We analyse how contributions to the public good evolve under different treatments, considering both the total tokens contributed and the share of endowment contributed. Figure 2 illustrates the average and median tokens contributed to the public good, segmented by treatment group.[9] Our dynamic PGG shows a significant deviation from static models, as evidenced by the general uptick in average contributions across all treatments except the *combined* treatment depicted in Figure 2(a). This rise contradicts previous literature on static PGGs, which typically show a decline in contributions over time.

The increase in contributions is most pronounced in the *punishment* treatment, suggesting that punishment mechanisms effectively sustain higher cooperation. However, this effect is not uniform across groups; a few outlier groups skew the average, while most groups exhibit lower contributions.[10] As depicted in Figure 2(b), the median contributions decrease across all groups, with the *punishment* and *combined* treatments experiencing a notable reduction, culminating in zero contributions in the final rounds. A clear gap emerges between the *baseline* and *imperfect monitoring* treatments compared to the *punishment* and *combined* treatments. This divergence between average and median contributions indicates a right-skewed distribution, where a small number of highly cooperative groups drive the mean

[9] Throughout the study, we observed significant outliers, especially in the *punishment* treatment, where a few particularly cooperative groups exponentially increased their wealth and contributions. These outliers skewed the results, prompting us to include both average and median contributions in our analysis.

[10] Table 2 reveals that the standard deviation of contributions (measured in tokens) is substantially higher in the punishment treatment compared to other treatments, as indicated by the values in parentheses. This disparity becomes even more pronounced in the final round (Round 10), with a standard deviation of 187 in Round 10, compared to 83.1, 62.9, and 13.3 in the *baseline*, *imperfect monitoring*, and *combined* treatments, respectively. Furthermore, Table B1 in Appendix B provides the standardized variance of group-level contributions for each treatment across all rounds, providing additional context.

upward, while the majority contribute less, below the average. Thus, while punishment raises average contributions, it amplifies disparities between groups.

Imperfect monitoring, when implemented alone, does not significantly differ from the *baseline* treatment in terms of average or median contributions, suggesting its limited impact on cooperative behaviour. In contrast, the combination of punishment and imperfect monitoring produces the lowest contribution rates across all treatments. This outcome likely stems from misdirected punishments, which undermine cooperation and discourage sustained contribution efforts.

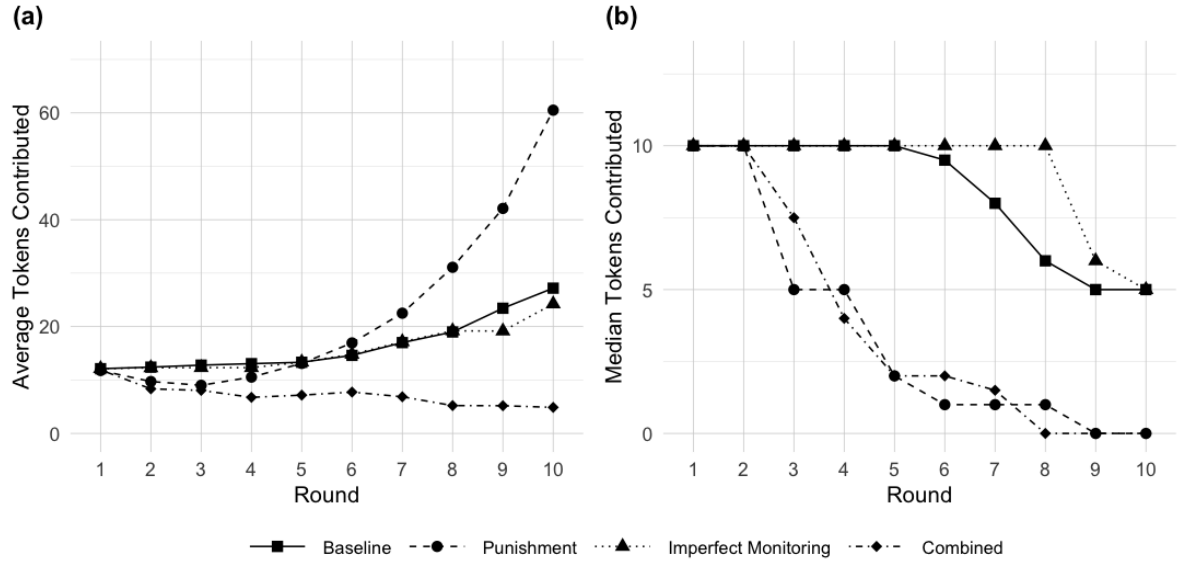

Figure 2 The average and median amount of tokens contributed to the public good over 10 rounds broken down by the four treatment groups. See Appendix A for the corresponding data table.

Next, we examine contributions to the public good as a share of one's endowment. We observe a gradual decline in contributions across all treatments (see Figure 3). Figure 3(a) shows that average contributions start at 55-60% of endowments in round 1, consistent with prior research. Typically, this percentage drops to around 10% in static games over multiple rounds (Chaudhuri, 2011; Ledyard, 1995). However, in our dynamic setting, contributions stabilize at a slightly higher 20% for all groups. This stabilization is expected, as dynamic incentives with potential exponential growth encourage higher contributions compared to static endowments. There are no visible differences between treatments in terms of average contribution share. Figure 3(b) shows median contribution shares, which mirror the average trend but at a lower level. Median relative contributions start at 50% and stabilize between 0 and 10%. This trend reflects the influence of highly cooperative outlier groups, which skew the group average to a higher level. Again, we find no visible differences between the treatments.

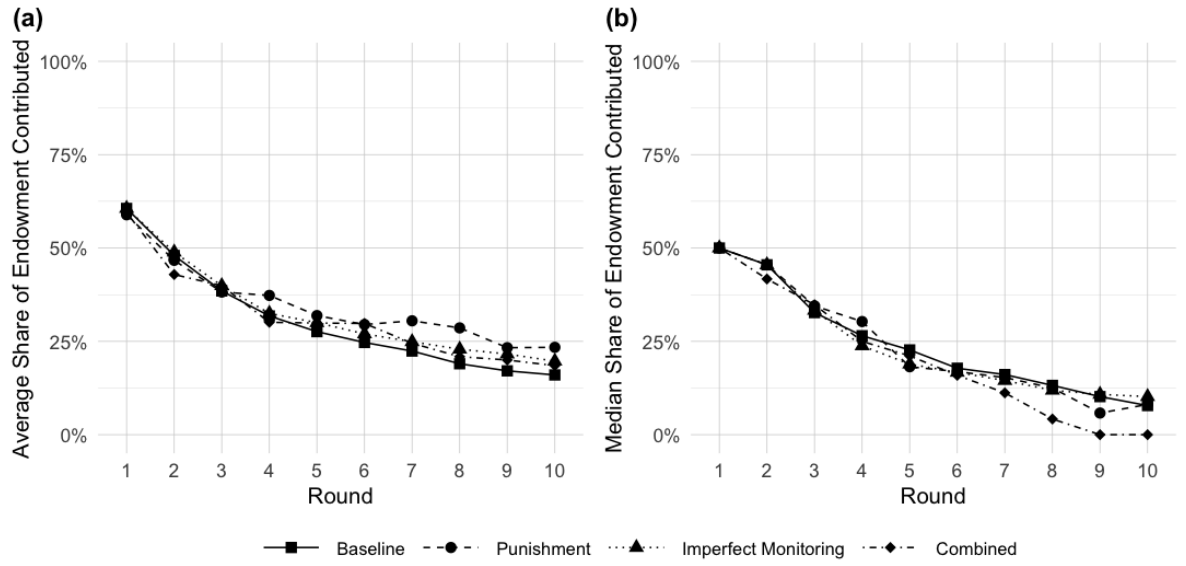

Figure 3 The average and median amount of tokens as a share of starting endowment contributed to the public good over 10 rounds broken down by the four treatment groups. See Appendix A for the corresponding data table.

Table 3 reports the average and median contribution rates (as a share of endowment) across all rounds, along with the p-values from two-sided Wilcoxon Rank Sum tests comparing the *baseline* treatment to the other treatments. These results reveal significant differences in contribution rates that were not evident in the corresponding figure: the *punishment* treatment differs substantially from the *baseline* (p-value < 0.001), as does the *combined* treatment (p-value < 0.001).

Table 3 Comparing contribution rate (as a share of endowment) between treatment groups across all rounds

	Treatment				Wilcoxon Rank Sum p-value		
	Baseline	**Punishment**	**Imperfect Monitoring**	**Combined**	**Punishment vs Baseline**	**Imperfect Monitoring vs Baseline**	**Combined vs Baseline**
Average Contribution Rate	30.56%	42.25%	32.80%	38.17%	<0.0001	0.844	0.0004
Median Contribution Rate	22.43%	34.48%	21.99 %	33.33%			
Standard Deviation	29.36	35.23	31.01	33.27			
Participants / Groups	120 / 40	111 / 37	105 / 35	114 / 38			

* Before conducting the Wilcoxon Rank Sum test, we averaged each participant's contribution rates across all rounds, reducing the dataset to one observation per participant to ensure independence of observations.

5.1.2. Wealth Accumulation

We next consider the varied effect of the different treatments on wealth. The wealth sums the endowment of all participants in a given group at the beginning of the following period. In other words,

wealth in round t will be the amount of tokens generated in round $t-1$. To ensure a comprehensive analysis, we report both the average and median wealth across the four treatments.

Figure 3(a) shows that wealth increases over rounds for all treatments except the *combined* treatment, where it stagnates around 19-24 tokens. While both the *punishment* and *combined* treatments begin with lower initial wealth, only the *punishment* treatment experiences a steep rise, largely driven by a few outlier groups, as discussed in Section 5.1. In contrast, Figure 3(b) shows the trends in median wealth. The *baseline* and *imperfect monitoring* treatments exhibit steady increases, whereas the *punishment* and *combined* treatments display declines over time. Notably, *imperfect monitoring* alone has a negligible effect on wealth accumulation. However, the combination of imperfect monitoring and punishment is associated with the lowest wealth accumulation across rounds.

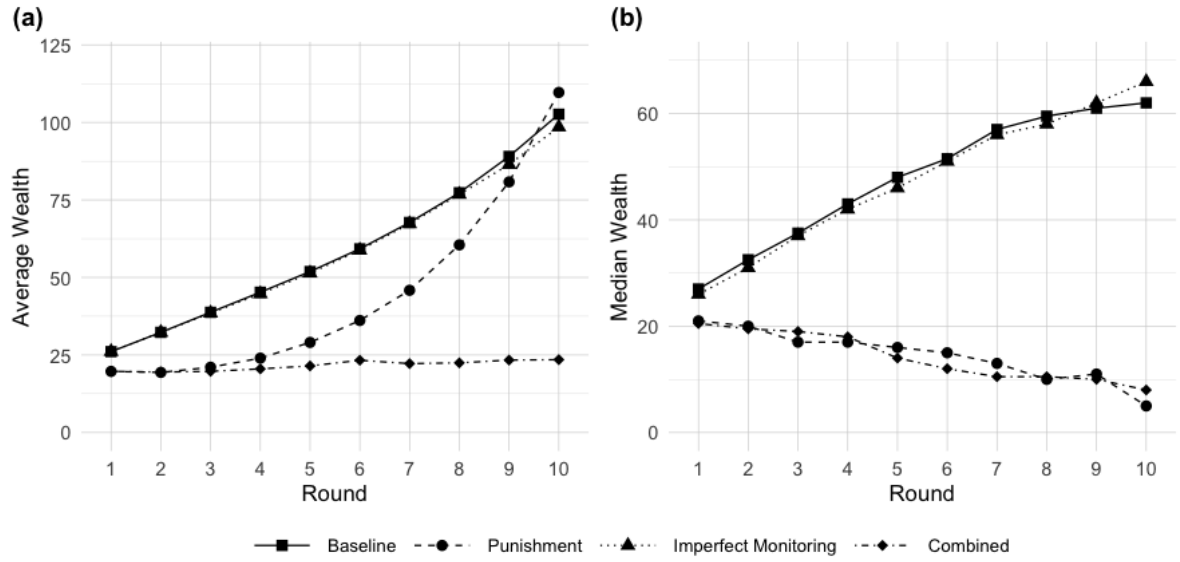

Figure 4 The average and median wealth generated over rounds broken down by the four treatment groups. See Appendix A for the corresponding data table.

We performed two-sided Wilcoxon Rank Sum tests to compare the final round wealth across our three treatments with the *baseline* condition, as shown in Table 3. The results, consistent with our graphical representations, indicated that wealth in both the *punishment* and *combined* treatments significantly differed from the *baseline* in round 10, whereas the *imperfect monitoring* treatment did not show a significant difference.

Table 4 Comparing accumulated wealth (in tokens) between treatment groups in round 10

	Treatment				Wilcoxon Rank Sum p-value		
	Baseline	**Punishment**	**Imperfect Monitoring**	**Combined**	**Punishment vs Baseline**	**Imperfect Monitoring vs Baseline**	**Combined vs Baseline**
Average Wealth	102.63	109.66	98.51	23.491	0.0002	0.633	<0.0001
Median Wealth	62	5	66	8			
Standard Deviation	130.84	248.58	111.65	38.18			
Participants / Groups	120 / 40	111 / 37	105 / 35	114 / 38			

5.1.3. Inequality

A central focus of our paper is to investigate how inequality is affected by our treatments. In order to measure inequality, we calculated the Gini coefficients for each three-member group. We removed the observations for groups where each member ended up with zero tokens. Next, in line with previous analyses, we report both the average and median Gini coefficients for each treatment. These results are presented in Figure 5.

The graphs reveal that inequality gradually rises for all groups as the rounds progress. In the *combined* treatment, both average and median inequality rise rapidly, peaking in the final round. Consistent with our previous analyses, we observe a gap between the average and median Gini coefficients in the *punishment* treatment. The average Gini coefficient suggests high inequality by the final rounds, while the median remains close to the baseline. Inequality in the *imperfect monitoring* treatment is moderately but consistently higher than in the baseline.

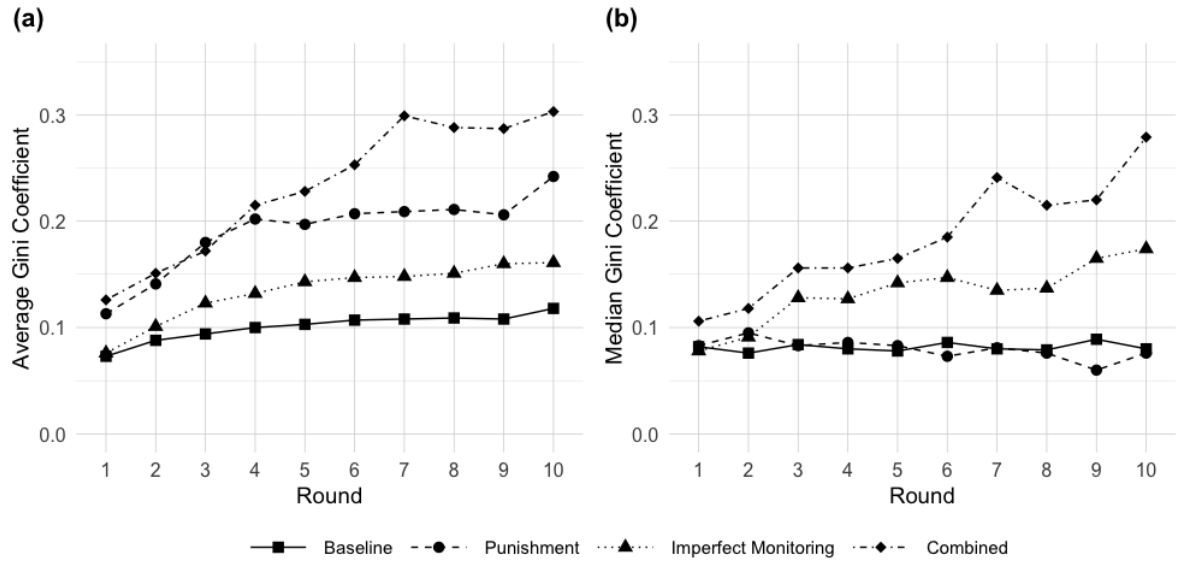

Figure 5 The average and median Gini coefficient over rounds broken down by the four treatment groups. See Appendix A for the corresponding data table.

We performed two-sided Wilcoxon Rank Sum tests to compare final round inequality across three treatments against the baseline condition, as shown in Table 5. The results indicate significant differences in final round inequality between the *imperfect monitoring* and *baseline* conditions, as well as between the *combined* treatment and *baseline*. However, we found no significant difference between the *punishment* and *baseline* conditions.

Table 5 Comparing inequality (Gini Coefficient) between treatment groups in round 10

	Treatment				Wilcoxon Rank Sum p-value		
	Baseline	**Punishment**	**Imperfect Monitoring**	**Combined**	**Punishment vs Baseline**	**Imperfect Monitoring vs Baseline**	**Combined vs Baseline**
Average Inequality	0.118	0.242	0.161	0.303	0.788	0.018	0.0005
Median Inequality	0.0804	0.0762	0.174	0.279			
Standard Deviation	0.0963	0.284	0.0882	0.226			
Participants / Groups	120 / 40	111 / 37	105 / 35	114 / 38			

5.1.4. Regression Analyses

We conducted linear mixed-effects regressions to examine the dynamics of group-level metrics, specifically the average contribution rate (as a share of endowment), total wealth, and the Gini coefficient. These regressions included three dummy-coded treatments and the round number as fixed effects, with the *baseline* treatment serving as the reference. To account for repeated measurements

from the same groups across the 10 rounds, we included group ID as a random effect. Table 6 summarizes these results, showing how contributions, wealth, and inequality evolve under varying experimental setups.

Our findings show that punishment significantly increases the average share of endowments that group members contribute, as seen in model (1). The interaction between the *punishment* treatment and round number in model (2) reveals a positive and significant effect, indicating that punishment becomes more effective in fostering contributions over time. However, there's also a small, consistent decline in contributions as rounds progress, regardless of treatment. This may reflect a natural tendency for cooperation to wane over time, especially in settings with a clear endpoint, consistent with Figure 3.

When it comes to total wealth, the *combined* treatment has a significant negative impact, an influence that grows stronger over time, as evidenced by the interaction term in model (2). Conversely, punishment alone does not significantly impact wealth, suggesting that while punishment enhances contribution rates, its cost dampens the net economic benefits. We also see that wealth tends to grow over time in all treatments. This pattern suggests that, even as cooperation declines slightly, the compounding effects of earlier contributions still support overall wealth growth.

The analysis of the Gini-coefficient highlights that both punishment and the combined treatment significantly increase inequality, with the combined treatment having a stronger effect. Over time, these treatments widen the inequality, as shown by the significant interaction between treatment and round number in model (2). Disregarding any treatment effect, inequality naturally rises as groups progress through the rounds, likely due to small differences in initial resources or strategies snowballing over time.

These results show the complex balance between encouraging cooperation, maintaining overall wealth, and keeping inequality in check. Punishment-based can successfully boost contributions, but they come with costs in the form of stagnating wealth and growing inequality. Moreover, when punishment is paired with imperfect monitoring, the negative effects become starker, resulting in diminishing cooperation, substantial wealth losses, and escalating inequalities.

Table 6 Results of linear mixed effects regression models to explain share of endowment contributed to the public good, accumulated wealth and Gini-coefficients

Regressors	Share of Endowment Contributed		Wealth Accumulated		Gini-coefficient	
	(1)	(2)	(1)	(2)	(1)	(2)
Fixed Effects						
Intercept	0.514 ***	0.409 ***	71.408 ***	-43.816	0.040 *	-0.014
	(0.030)	(0.117)	(23.335)	(92.285)	(0.021)	(0.081)
Punishment	0.086 **	0.012	-43.222	-57.213	0.092 ***	0.055 *
	(0.043)	(0.047)	(31.831)	(39.827)	(0.029)	(0.032)
Imperfect Monitoring	0.022	0	-2.831	3.839	0.033	0.014
	(0.042)	(0.046)	(32.300)	(39.896)	(0.029)	(0.032)
Combined	0.062	-0.017	-112.328 ***	10.227	0.139 ***	0.052
	(0.042)	(0.046)	(31.613)	(39.373)	(0.029)	(0.032)
Round	-0.038 ***	-0.046 ***	19.210 ***	25.159 ***	0.011 ***	0.003 **
	(0.001)	(0.002)	(1.380)	(2.697)	(0.001)	(0.001)
Round x Punishment		0.019 ***		3.921		0.008 ***
		(0.004)		(3.922)		(0.002)
Round x Imperfect Monitoring		0.006 *		-0.994		0.005 **
		(0.003)		(3.952)		(0.002)
Round x Combined		0.017 ***		-23.277 ***		0.020 ***
		(0.004)		(3.893)		(0.002)
Control for Demographics^ and Risk Aversion	No	Yes	No	Yes	No	Yes
Random Effects (Variance)						
Group ID	0.03	0.03	17116.86	17869.89	0.02	0.02
Observations [a]	1,296	1,240	1,500	1,500	1367	1367
Participants / Groups	450 / 150	432 / 144	450 / 150	432 / 144	450 / 150	432 / 144
AIC	-1143.866	-1061.452	19649.931	18841.601	-2325.959	-2238.365
Marginal R^2 / Conditional R^2	0.219 / 0.728	0.246 / 0.737	0.112 / 0.485	0.139 / 0.512	0.141 / 0.719	0.206 / 0.747

*** p <0.01, ** p < 0.05, * p <0.1; ^ Demographic variables include average age and percentage of females within each group.
Note: Figures in parentheses are standard errors
[a] We collected data from 150 groups, with each group providing 10 observations corresponding to each round. However, we excluded observations where all members ended up with zero earnings, as this made it impossible to calculate the share of endowment contributed as well as Gini coefficient.

5.2. Anatomy of Punishment

This section examines how participants administer punishment over successive rounds in the *punish* and *combined* treatments. Figure 6 illustrates the average punishment cost (in tokens) borne by

participants.[11] The figure reveals that round 3 onward, participants exhibit increased punitiveness in the *combined* treatment, with a sharp spike in round 7. Wilcoxon Rank Sum tests confirm that participants in the combined treatment incur significantly higher punishment costs (p-value < 0.05, see Table 7). Furthermore, consistent with the declining contributions observed across rounds in both treatments, the tokens spent on punishment generally decrease over time, except for the notable surge in round 7 in the *combined* treatment.

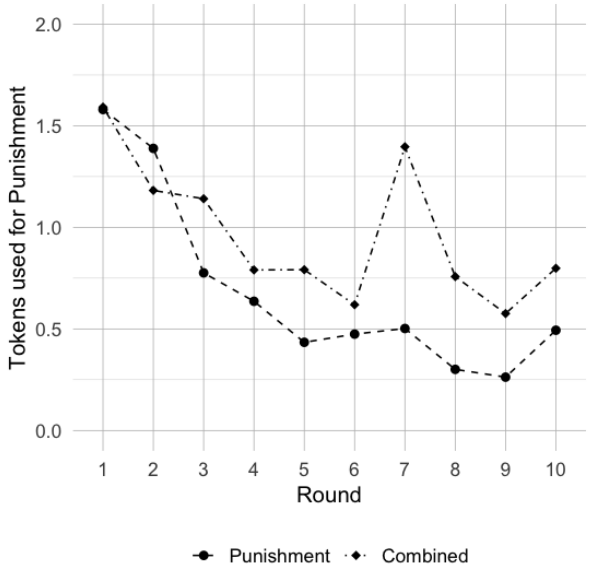

Figure 6 Cost of punishment over rounds broken down by the two treatment groups with peer punishment.

To better understand the nature of punishment in these treatments, we categorize punishment into pro-social and anti-social. In the voluntary contribution mechanism (VCM) literature, punishment is deemed pro-social if a player punishes another who contributed a smaller share of their endowment to the public good than themselves. Conversely, punishment is considered anti-social if it is directed at a player who contributed a larger share.[12]

Figure 7 disaggregates punishments into pro-social and anti-social categories. Regarding pro-social punishment, no distinct trend emerges between the treatments, even though, on average, participants in the combined treatment bear a higher cost for prosocial punishments (Wilcoxon Rank Sum test p-value < 0.05, see Table 7). In terms of anti-social punishment, the trends are evident: round 3 onward, participants in the combined treatment allocate significantly more tokens to inflict antisocial punishments (Wilcoxon Rank Sum test p-value < 0.05, see Table 7). It looks like the higher overall

[11] It is important to clarify that this cost refers to the expense incurred by those administering punishment, not the tokens deducted from those punished (which are three times the cost).
[12] Additionally, there exists a third category of punishment—when a player punishes another who contributed an equal share—which does not fall under prosocial or antisocial categories.

punishment observed in the combined treatment is predominantly driven by antisocial punishments, including the round 7 surge. This trend is likely attributable to the effects of imperfect monitoring inherent in the combined treatment leading to greater dissemination of misdirected punishments.

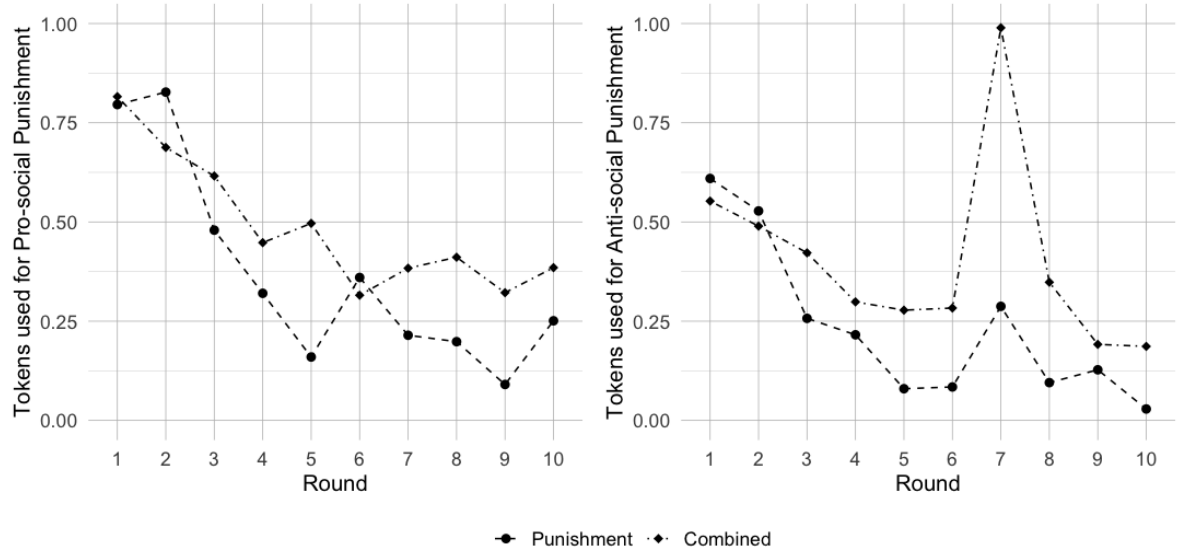

Figure 7 Cost of pro-social and anti-social punishment over rounds broken down by the two treatment groups with peer punishment.

Table 7 Comparing costs of punishment between *punishment* and *combined* treatments across all rounds

	Treatments		Wilcoxon Rank Sum p-value*
	Punishment	**Combined**	**Punishment vs Combined**
Average Punishment Cost	0.727	0.988	0.0419
Standard Deviation	1.831	2.770	
Average Pro-social Punishment Cost	0.334	0.427	0.0248
Standard Deviation	1.297	1.335	
Average Anti-social Punishment Cost	0.211	0.350	0.0158
Standard Deviation	0.888	2.149	
Participants / Groups	111 / 37	114 / 38	

* Before conducting the Wilcoxon Rank Sum test, we averaged each participant's punishment costs across all rounds, reducing the dataset to one observation per participant to ensure independence of observations.

5.2.1. Regression Analyses

We conducted a series of linear mixed-effects regressions to examine the costs associated with punishment, including pro-social and anti-social punishments. Since punishments were implemented exclusively in the *punishment* and *combined* treatments, our analysis is restricted to data from these two treatment. Aside from the treatment scope, the model specifications are identical to those reported in Table 6.

Regarding the overall cost of punishment, the *combined* treatment exhibits a marginally significant positive effect in model (1). However, this effect loses statistical significance when

demographic and risk aversion controls are included in model (2). For the cost of pro-social punishment, the *combined* treatment does not exert a significant effect in either model. In contrast, the cost of anti-social punishment shows a marginally significant positive effect in model (1), which dissipates when controls are added in model (2). Moreover, the interaction between the *combined* treatment and round is not significant across both models. Consistently across all punishment types, the round variable is negative and statistically significant, reflecting a decline in punishment costs over time. This trend reflects a reduction in either the opportunities for or the willingness to engage in punishment as group dynamics stabilize.

Overall, these results demonstrate that the *combined* treatment has a minimal impact on punishment costs. The absence of significant interaction effects between the *combined* treatment and round further suggests that the observed declining trends in punishment costs are predominantly driven by temporal dynamics rather than treatment-specific factors.

Table 8 Results of linear mixed effects regression models to explain the costs of punishment

Regressors	**Cost of Punishment**		**Cost of Pro-social Punishment**		**Cost of Anti-social Punishment**	
	(1)	**(2)**	**(1)**	**(2)**	**(1)**	**(2)**
Fixed Effects						
Intercept	1.254 ***	1.156 **	0.696 ***	0.792 ***	0.443 ***	0.318
	(0.147)	(0.566)	(0.067)	(0.234)	(0.086)	(0.300)
Combined	0.265 *	-0.046	0.094	0.003	0.139 *	-0.052
	(0.155)	(0.265)	(0.061)	(0.126)	(0.076)	(0.162)
Round	-0.090 ***	-0.122 ***	-0.066 ***	-0.076 ***	-0.042 ***	-0.058 ***
	(0.019)	(0.029)	(0.009)	(0.014)	(0.012)	(0.018)
Round x Combined		0.065		0.021		0.03
		(0.041)		(0.020)		(0.025)
Control for Demographics^ and Risk Aversion	No	Yes	No	Yes	No	Yes
Random Effects (Variance)						
Group ID	0.21	0.20	0.02	0.01	0.02	0.02
Observations [a]	637	599	637	599	637	599
Participants / Groups	225 / 75	213 / 71	225 / 75	213 / 71	225 / 75	213 / 71
AIC	2288.753	2214.234	1527.905	1499.542	1867.084	1824.421
Marginal R^2 / Conditional R^2	0.038 / 0.133	0.048 / 0.133	0.042 / 0.169	0.054 / 0.169	0.016 / 0.047	0.020 / 0.049

*** p <0.01, ** p < 0.05, * p <0.1; ^ Demographic variables include average age and percentage of females within each group.
Note: Figures in parentheses are standard errors

[a] We collected data from 150 groups, with each group providing 10 observations corresponding to each round. However, we excluded observations where all members ended up with zero earnings, as this made it impossible to calculate the Gini coefficient. This exclusion resulted in a total of 1,367 observations.

6. Conclusion

Our findings underscore the complexity of fostering cooperation in social dilemmas, highlighting the intricate trade-offs between encouraging collective action, sustaining overall wealth, and maintaining economic equity. The narrative that emerges is one of paradox: mechanisms designed to enforce cooperation often sow the seeds of discord, amplifying inequality and eroding the very benefits they aim to secure.

Punishment emerges as a double-edged sword. While it elevates average contributions to the public good, this increase masks a troubling undercurrent—cooperation becomes concentrated among a few highly engaged groups, leaving others behind. The result is a growing divide, where outliers drive the averages upward, yet the majority experience stagnation or decline. This dynamic not only underscores the limitations of punishment as a universal tool for fostering cooperation but also exposes its potential to exacerbate disparities among participants. The introduction of imperfect monitoring complicates this narrative further. On its own, it fails to significantly alter behaviour, largely mirroring the baseline dynamics. Yet, when paired with punishment, the effects are stark. The use of punishments proliferate, cooperation collapses, and wealth accumulation grinds to a halt. This combination not only undermines the cooperative gains achieved through punishment alone but also accelerates the rise of inequality. Anti-social punishments, disproportionately observed in the *combined* treatment, redirect valuable resources away from productive efforts and into cycles of retribution. Pro-social punishments, though present, fail to counterbalance this trend, further entrenching disparities and discouraging sustained cooperation.

The broader narrative of these findings highlights the complex interplay between punishment mechanisms, the accuracy of public information, and the behavioural responses they elicit. Punishment can be effective on its own to increase cooperation, but punishment without precision risks a wide array of negative outcomes. These findings carry significant implications for policy design, particularly in contexts where public goods are provided under conditions of uncertainty. While punishment can be a powerful tool for enhancing cooperation, its application must be carefully calibrated to avoid unintended consequences. Policymakers should be wary of introducing punitive measures in settings where monitoring is imperfect, as this may inadvertently harm collective welfare and exacerbate inequalities.

In conclusion, our study contributes to the broader understanding of the dynamics of cooperation, punishment, and inequality. It underscores the importance of considering the quality of information and the potential for misdirected punitive actions when designing interventions aimed at promoting public good provision.

7. References

Acemoglu, D., & Jackson, M. O. (2015). History, Expectations, and Leadership in the Evolution of Social Norms. *The Review of Economic Studies*, *82*(2 (291)), 423–456. http://www.jstor.org/stable/43551536

Altemeyer, B. (1981). *Right-wing Authoritarianism*. University of Manitoba Press.

Ambrus, A., & Greiner, B. (2012). Imperfect Public Monitoring with Costly Punishment: An Experimental Study. *American Economic Review*, *102*(7), 3317–3332. https://doi.org/10.1257/aer.102.7.3317

Ambrus, A., & Greiner, B. (2019). Individual, Dictator, and Democratic punishment in public good games with perfect and imperfect observability. *Journal of Public Economics*, *178*, 104053. https://doi.org/10.1016/j.jpubeco.2019.104053

Andreoni, J., & Gee, L. K. (2012). Gun for hire: Delegated enforcement and peer punishment in public goods provision. *Journal of Public Economics*, *96*(11–12), 1036–1046. https://doi.org/10.1016/j.jpubeco.2012.08.003

Axelrod, R. (1984). *The evolution of cooperation*. Basic Books.

Cadigan, J., Wayland, P. T., Schmitt, P., & Swope, K. (2011). An experimental dynamic public goods game with carryover. *Journal of Economic Behavior & Organization*, *80*(3), 523–531. https://doi.org/10.1016/j.jebo.2011.05.010

Chaudhuri, A. (2011). Sustaining cooperation in laboratory public goods experiments: a selective survey of the literature. *Experimental Economics*, *14*(1), 47–83. https://doi.org/10.1007/s10683-010-9257-1

Chen, D. L., Schonger, M., & Wickens, C. (2016). oTree—An open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance*, *9*, 88–97. https://doi.org/10.1016/j.jbef.2015.12.001

Coleman, J. S. (1988). Social Capital in the Creation of Human Capital. *American Journal of Sociology*, *94*, S95–S120. http://www.jstor.org/stable/2780243

Fehr, E., & Gächter, S. (2000). Cooperation and Punishment in Public Goods Experiments. *American Economic Review*, *90*(4), 980–994. https://doi.org/10.1257/aer.90.4.980

Fehr, E., & Gächter, S. (2002). Altruistic punishment in humans. *Nature*, *415*(6868), 137–140. https://doi.org/10.1038/415137a

Fischbacher, U., Gächter, S., & Fehr, E. (2001). Are people conditionally cooperative? Evidence from a public goods experiment. *Economics Letters*, *71*(3), 397–404. https://doi.org/10.1016/S0165-1765(01)00394-9

Gächter, S., Mengel, F., Tsakas, E., & Vostroknutov, A. (2017). Growth and inequality in public good provision. *Journal of Public Economics*, *150*, 1–13. https://doi.org/10.1016/j.jpubeco.2017.03.002

Gächter, S., Renner, E., & Sefton, M. (2008). The Long-Run Benefits of Punishment. *Science*, *322*(5907), 1510–1510. https://doi.org/10.1126/science.1164744

Gintis, H. (2000). Strong Reciprocity and Human Sociality. *Journal of Theoretical Biology*, *206*(2), 169–179. https://doi.org/10.1006/jtbi.2000.2111

Grechenig, K., Nicklisch, A., & Thöni, C. (2010). Punishment Despite Reasonable Doubt—A Public Goods Experiment with Sanctions Under Uncertainty. *Journal of Empirical Legal Studies*, *7*(4), 847–867. https://doi.org/10.1111/j.1740-1461.2010.01197.x

Henrich, J., & Muthukrishna, M. (2021). The Origins and Psychology of Human Cooperation. *Annual Review of Psychology*, *72*(1), 207–240. https://doi.org/10.1146/annurev-psych-081920-042106

Hill, K., Kaplan, H., & Hawkes, K. (1993). On Why Male Foragers Hunt and Share Food. *Current Anthropology*, *34*(5), 701–710. http://www.jstor.org/stable/2744280

Ledyard, J. O. (1995). 2. Public Goods: A Survey of Experimental Research. In J. H. Kagel & A. E. Roth (Eds.), *The Handbook of Experimental Economics* (pp. 111–194). Princeton University Press. https://doi.org/10.1515/9780691213255-004

Markussen, T., Putterman, L., & Tyran, J.-R. (2016). Judicial error and cooperation. *European Economic Review*, *89*, 372–388. https://doi.org/10.1016/j.euroecorev.2016.08.004

Marwell, G., & Ames, R. E. (1979). Experiments on the Provision of Public Goods. I. Resources, Interest, Group Size, and the Free-Rider Problem. *American Journal of Sociology*, *84*(6), 1335–1360. https://doi.org/10.1086/226937

Mengel, F., Weidenholzer, S., & Mohlin, E. (2021). Collective Incentives and Cooperation with Imperfect Monitoring. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3823041

Milinski, M., Semmann, D., & Krambeck, H.-J. (2002). Reputation helps solve the 'tragedy of the commons.' *Nature*, *415*(6870), 424–426. https://doi.org/10.1038/415424a

Noussair, C., & Soo, C. (2008). Voluntary contributions to a dynamic public good: Experimental evidence. *Economics Letters*, *98*(1), 71–77. https://doi.org/10.1016/j.econlet.2007.04.008

Nowak, M. A. (2006). Five Rules for the Evolution of Cooperation. *Science*, *314*(5805), 1560–1563. https://doi.org/10.1126/science.1133755

Ostrom, E. (1998). A Behavioral Approach to the Rational Choice Theory of Collective Action: Presidential Address, American Political Science Association, 1997. *American Political Science Review*, *92*(1), 1–22. https://doi.org/DOI: 10.2307/2585925

Ostrom, E. (2015). *Governing the Commons*. Cambridge University Press. https://doi.org/10.1017/CBO9781316423936

Ostrom, E., Walker, J., & Gardner, R. (1992). Covenants with and without a Sword: Self-Governance Is Possible. *American Political Science Review*, *86*(2), 404–417. https://doi.org/DOI: 10.2307/1964229

Pratto, F., Sidanius, J., Stallworth, L. M., & Malle, B. F. (1994). Social dominance orientation: A personality variable predicting social and political attitudes. *Journal of Personality and Social Psychology*, *67*(4), 741–763. https://doi.org/10.1037/0022-3514.67.4.741

Rand, D. G., & Nowak, M. A. (2013). Human cooperation. *Trends in Cognitive Sciences*, *17*(8), 413–425. https://doi.org/10.1016/j.tics.2013.06.003

Rockenbach, B., & Wolff, I. (2019). The Dose Does it: Punishment and Cooperation in Dynamic Public-Good Games. *Review of Behavioral Economics*, *6*(1), 19–37.

Sadrieh, A., & Verbon, H. A. A. (2006). Inequality, cooperation, and growth: An experimental study. *European Economic Review*, *50*(5), 1197–1222. https://doi.org/10.1016/j.euroecorev.2005.01.009

Sefton, M., Shupp, R., & Walker, J. M. (2007). The Effect Of Rewards And Sanctions In Provision Of Public Goods. *Economic Inquiry*, *45*(4), 671–690. https://doi.org/10.1111/j.1465-7295.2007.00051.x

Stibbard-Hawkes, D. N. E., Smith, K., & Apicella, C. L. (2022). Why hunt? Why gather? Why share? Hadza assessments of foraging and food-sharing motive. *Evolution and Human Behavior*, *43*(3), 257–272. https://doi.org/https://doi.org/10.1016/j.evolhumbehav.2022.03.001

West, S. A., Griffin, A. S., & Gardner, A. (2007). Social semantics: altruism, cooperation, mutualism, strong reciprocity and group selection. *Journal of Evolutionary Biology*, *20*(2), 415–432. https://doi.org/10.1111/j.1420-9101.2006.01258.x

Appendix A: Summary Statistics

Average Tokens Contributed

Treatment / Round	1	2	3	4	5	6	7	8	9	10
Baseline	12.117	12.408	12.783	13.058	13.325	14.608	17	18.992	23.433	27.192
Punishment	11.775	9.712	9.009	10.541	13.108	16.937	22.505	31.099	42.117	60.505
Imperfect Monitoring	12.124	12.286	12.314	12.305	13.352	14.81	17.162	19.2	19.162	24.238
Combined	11.895	8.342	8.053	6.754	7.167	7.737	6.86	5.211	5.184	4.877
All	11.978	10.684	10.544	10.664	11.718	13.489	15.827	18.536	22.422	29.067

Median Tokens Contributed

Treatment / Round	1	2	3	4	5	6	7	8	9	10
Baseline	10	10	10	10	10	9.5	8	6	5	5
Punishment	10	10	5	5	2	1	1	1	0	0
Imperfect Monitoring	10	10	10	10	10	10	10	10	6	5
Combined	10	10	7.5	4	2	2	1.5	0	0	0
All	10	10	10	8	6	5	5	4	3	2

Average Tokens as a Share of Endowment Contributed

Treatment / Round	1	2	3	4	5	6	7	8	9	10
Baseline	0.606	0.48	0.385	0.318	0.276	0.247	0.224	0.19	0.171	0.16
Punishment	0.589	0.467	0.382	0.373	0.319	0.295	0.305	0.286	0.233	0.234
Imperfect Monitoring	0.606	0.489	0.399	0.325	0.3	0.27	0.249	0.229	0.216	0.197
Combined	0.595	0.429	0.399	0.301	0.299	0.298	0.245	0.209	0.2	0.185
All	0.599	0.466	0.391	0.328	0.297	0.276	0.254	0.227	0.203	0.192

Average Wealth

Treatment / Round	1	2	3	4	5	6	7	8	9	10
Baseline	26.08	32.3	38.77	45.23	51.91	59.23	67.73	77.31	89.03	102.63
Punishment	19.71	19.39	21.05	23.98	29.05	36.13	45.86	60.50	80.81	109.66
Imperfect Monitoring	26.11	32.28	38.51	44.71	51.39	58.78	67.31	76.79	86.4	98.51
Combined	19.728	19.5	19.728	20.518	21.474	23.298	22.219	22.465	23.36	23.491
All	22.907	25.867	29.513	33.609	38.44	44.322	50.704	59.147	69.751	83.353

Median Wealth

Treatment / Round	1	2	3	4	5	6	7	8	9	10
Baseline	27	32.5	37.5	43	48	51.5	57	59.5	61	62
Punishment	21	20	17	17	16	15	13	10	11	5
Imperfect Monitoring	26	31	37	42	46	51	56	58	62	66
Combined	20.5	19.5	19	18	14	12	10.5	10.5	10	8
All	24	28	30	34	37.5	39.5	39	41	44	46

Average Gini Coefficient

Treatment / Round	1	2	3	4	5	6	7	8	9	10
Baseline	0.074	0.088	0.094	0.1	0.103	0.107	0.108	0.109	0.108	0.118
Punishment	0.113	0.141	0.18	0.202	0.197	0.207	0.209	0.211	0.206	0.242
Imperfect Monitoring	0.076	0.101	0.123	0.132	0.143	0.147	0.148	0.151	0.16	0.161
Combined	0.126	0.151	0.172	0.215	0.228	0.253	0.299	0.288	0.287	0.303
All	0.097	0.119	0.140	0.158	0.162	0.172	0.183	0.181	0.181	0.195

Median Gini Coefficient

Treatment / Round	1	2	3	4	5	6	7	8	9	10
Baseline	0.082	0.076	0.085	0.079	0.077	0.086	0.08	0.079	0.089	0.08
Punishment	0.083	0.095	0.083	0.086	0.083	0.073	0.081	0.076	0.06	0.076
Imperfect Monitoring	0.078	0.091	0.128	0.127	0.142	0.147	0.135	0.137	0.165	0.174
Combined	0.106	0.118	0.156	0.155	0.165	0.185	0.241	0.215	0.219	0.279
All	0.083	0.095	0.110	0.115	0.108	0.121	0.133	0.128	0.136	0.148

Appendix B: Other Tables and Figures

Standardized Variance of Tokens Contributed

Table B1

Treatment / Round	1	2	3	4	5	6	7	8	9	10
Baseline	2.690	7.859	17.774	27.407	46.748	90.191	155.994	285.360	448.954	762.490
Punishment	2.940	11.819	32.072	63.152	116.140	192.669	328.065	533.331	860.391	1324.333
Imperfect Monitoring	2.442	5.210	12.851	26.867	50.572	96.103	160.448	256.446	202.603	489.577
Combined	1.112	7.137	12.142	24.648	34.747	53.734	67.252	64.753	71.627	84.266