Journal of Behavioral Economics for Policy, Vol. 7, Issue 2, 11-20, 2023

Weak and Strong Formal Institutions in Resolving Social Dilemmas: Are They Double-Edged Swords?

Rati Mekvabishvili [1]*

Abstract
Many modern societies sustain large-scale cooperation among strangers and maintain the provision of public goods through well-functioning top-down formal institutions. However, it is important to understand the differences between weak and strong formal institutions in achieving two key goals in social dilemma situations: sustaining socially beneficial equilibria and fostering individual prosocial behavior. Additionally, we need to examine what happens to cooperation when the credibility of a formal institution is undermined and what occurs when it ceases to function. In this novel experiment of a repeated public goods game, we explore the effects of an exogenous centralized punishment mechanism with a low probability, which serves as a weak formal institution, and compare it with a strong formal institution. Our findings are encouraging, as they demonstrate that even under a weak formal institution, relatively high levels of cooperation can be sustained. However, irrespective of whether the punishment probability for free riders is low or high, once the punishment mechanism is removed, cooperation breaks down to a similarly low level. This suggests that regardless of the strength of the formal institution, there is an alike effect of crowding out an individual's intrinsic motivation for cooperation. Therefore, the application of a centralized punishment mechanism as a policy tool to promote cooperation, regardless of its strength, appears to be a double-edged sword: socially beneficial outcome and intrinsically motivated cooperation hardly can be attained simultaneously.

JEL Classification: C9; H41; D02

Keywords
formal institutions — public good — centralized punishment — crowding out — cooperation

[1] *Ivane Javakhishvili Tbilisi State University, Faculty of Economics and Business*
***Corresponding author**: rati.mekvabishvili@eab.tsu.edu.ge

Introduction

In society, formal institutions resolve many centrally important free-rider problems that are essential for the successful provision of public goods. A strong state is characterized by trustworthy and well-functioning state institutions, whose officials are honest, incorrupt, and effective in responding to the citizens' needs. However, fair and trustful state institutions can be considered second-order public goods, making them susceptible to free-riding and opportunistic behavior. Additionally, it is important to recognize that formal institutions often have limited scope. For instance, in tax enforcement, certain forms of income tax rely heavily on voluntary reporting. Similarly, in environmental contexts, enforcing anti-littering behavior can be challenging. Nevertheless, it is possible for norms established by institutions to "carry over" into the future and influence subsequent behavior in environments where those institutions may not directly apply. Therefore, understanding the degree to which institutions foster or discourage voluntary compliance with rules in the absence of formal enforcement is crucial for policymakers. A substantial and expanding body of evidence suggests that the impact of incentives depends on how they are designed and how they interact with intrinsic motivations.

Our paper contributes to the literature on studies of cooperative behavior in social dilemma situations with two extensions. Firstly, we explore the effectiveness of weak formal institutions in promoting cooperation within a single domain, and whether they have any crowding out[1] or spillover effects beyond their reach. Specifically, we conduct an experimental investigation using a weak exogenously imposed centralized punishment mechanism in a repeated Public Goods Game (PGG). Our primary focus is to understand the impact of this exogenous centralized punishment mechanism on cooperative behavior, both during its active presence and in subsequent behavior after its removal. To achieve this, the PGG experiment consists of two stages. First, we examine behavior in the presence of the centralized punishment mechanism (CPM). Then, we analyze behavior in subsequent rounds after the CPM has been removed. Furthermore, we introduce a novel mechanism of centralized punishment by differentiating between inspection and punishment. Specifically, participants may be subject to inspection without necessarily facing penalties.

[1] As for the definition and modelling of crowding out of intrinsic motivation, se Benabou and Tirole (2006). As for its discussion as a behavioral anomaly, see Frey (2017) and the literature review therein.

Secondly, our study extends our recent experiment Mekvabishvili (2021a), where in public goods game they examine cooperative behavior in presence and in absence of strong exogenous centralized punishment mechanism. Mekvabishvili (2021a) found that exposure to strong formal institutions that provide top-down motivation for cooperation substantially improve cooperation in their presence, but do not lead to more prosociality after their absence. The external incentive led to crowding out effect of the internal incentives to cooperate. In the same experimental setting, we now investigate the impact of a weak formal institution. This allows us to draw conclusions by comparing the levels of cooperation under weak and strong formal institutions within a single domain, i.e. the public goods game. These comparisons are important for distinguishing the effects of top-down weak and strong formal institutions on cooperative behavior in the specific domain.

The questions that motivate our study are as follows: How does a top-down weak formal institution promote cooperation during its direct exposure, and does this cooperation carry over into the future when the institution is absent within the same domain context? Alternatively, is the external incentive induced by the weak institution crowding out individuals' intrinsic motivation to cooperate, and if so, to what extent? The answers to these questions may have implications for the broader question of how exogenously imposed institutions with varying strength of incentives impact not only immediate behavior but also intrinsic motivations for cooperation and subsequent behavior in the future.

Our findings reveal that even a relatively low probability of centralized punishment can sustain high levels of cooperation in the public goods game. However, regardless of whether the punishment probability for free riders is low or high, once the centralized punishment mechanism is removed, cooperation collapses to a similarly low level. Thus, within the context of a single domain, both weak and strong formal institutions lead to a similar crowding out effect on individuals' intrinsic motivation for cooperation.

The remainder of the paper is organized as follows: Section 2 provides a review of the related literature; Section 3 describes the experimental design and procedures; Section 4 presents the experimental results of our investigation; Section 5 discusses the findings; Section 6 concludes the paper.

Related Literature

Our study relates to several strands of the literature. Firstly, there has been extensive research conducted over the past two decades on the impact of institutions in the provision of public goods. A seminal experimental study by Fehr and Gächter (2000) demonstrated that individuals are willing to punish free-riders, and that peer punishment enhances cooperation. Furthermore, they found that in the absence of peer punishment, cooperation tends to break down. However, one challenging aspect associated with peer punishment mechanisms is the potential for some players to misuse sanctioning incentives and undermine cooperation. For example, several experiments in public goods games with peer punishment have documented the existence of "antisocial" punishment, where sanctions are extensively used against cooperators rather than free-riders (Herrmann et al., 2008; Nikiforakis, 2008).

Another line of experimental studies has explored the effectiveness of endogenous centralized punishment, where one group member serves as a monitoring entity. Baldassarri and Grossman (2011) and O'Gorman et al. (2009) found that endogenous centralized punishment mechanisms can effectively promote cooperation. Putterman et al. (2011) presented a novel experimental study focusing on the design of sanction schemes. In their experiment, participants voted on whether to penalize group members and had to construct their own sanction scheme through voting on simple components. Remarkably, despite the absence of suggestive instructions and communication opportunities, the majority of groups selected a fully efficient regime within two or three votes. Additionally, Tyran and Feld (2006) conducted an experiment comparing the effects of endogenously and exogenously introduced 'mild' sanctions in a public goods game. In the endogenous treatment, subjects voted on whether to implement the sanction. The authors demonstrated that endogenous sanctions were more effective in increasing contributions compared to exogenously implemented sanctions.

Peysakhovich and Rand (2016) conducted an experimental study to examine the relationship between peer-based reputational incentives for cooperation and subsequent prosocial behavior. In the first stage, participants engaged in a series of repeated prisoner's dilemma games, and in the second stage, they played one-shot dictator games involving cooperation. The study found that the duration of repeated prisoner's dilemma games (leading to high versus low levels of bilateral cooperation) influenced subsequent giving in the dictator games, as well as other one-shot cooperation games. These results suggest that the norms of cooperation carry over into atypical situations that are beyond the reach of the institution promoting cooperation.

Stagnaro et al. (2017) extended the findings of Peysakhovich and Rand (2016) by applying repeated interactions not only between pairs of individuals but also involving formal top-down institutional punishment and group cooperation among more than two people. They manipulated institutional quality in a repeated PGG with an exogenously imposed centralized punishment institution. In the first stage, subjects played a ten-round PGG with an exogenous centralized punishment mechanism, and in the second stage, they played a one-shot dictator game (DG). The study revealed that the presence of centralized punishment led to significantly more prosocial behavior in the subsequent dictator game, providing evidence that the quality of institutions that individuals are exposed to in one domain "spills over" to subsequent prosocial behavior in another domain. In a recent PGG experiment conducted by Engel et al. (2021), the focus was on examining how the presence and nature of exogenously and endogenously imposed institutions that enforce prosocial

behavior in one domain affect behavior in another domain. The study found clear evidence supporting positive spillover effects between the domains.

However, it is important to consider the potential effect of external incentives, which may act as substitutes rather than complements to intrinsic motivations and subsequent cooperative behavior. Another area of research examines the "crowding-out effect" in economics, where an external incentive displaces intrinsic individual motivations for cooperation. Behavioral economists have cautioned that incentives can backfire by crowding out intrinsic motivation, especially when they are imposed from the top-down and perceived as controlling by individuals (Bowles and Polania-Reyes, 2012; Gneezy et al., 2011; Frey and Jegen, 2001; Gneezy and Rustichini, 2000). The crowding-out effect resulting from externally imposed incentives was observed in an earlier experimental study by Frohlich and Oppenheimer (2003).

Frohlich and Oppenheimer conducted an experiment to explore the effects of an incentive-compatible device (ICD) as an external incentive for cooperation in a repeated, linear, 5-person prisoner's dilemma game. They aimed to answer the main question of whether the behavioral effects of playing under an ICD carry over into the future. The study found that the ICD was successful in overcoming the tendency to free-ride, but when it was removed, a significantly lower level of cooperation was observed. Thus, no positive spillover effect of the ICD was observed. The researchers concluded that achieving the dual goals of collective welfare and fostering individual cooperative behavior simultaneously through the use of an ICD could be challenging.

When explicit incentives are applied to induce behavior change, such as increasing contributions to public goods, a potential conflict arises between the direct extrinsic effect of the incentives and their potential to crowd out intrinsic motivations in the short and long term. The fact that external incentives, such as punishment, often function more as messages than as effective incentives pose a challenge for policy designers (Gneezy and Rustichini, 2000; Funk, 2007; Galbiati and Vertova, 2014). For instance, in a well-known study conducted by Gneezy and Rustichini (2000), the authors sought to address the following problem: parents at a day-care center were frequently arriving late to pick up their children, causing a teacher to remain after closing time. In their field study, the researchers introduced a monetary fine for late-coming parents. Surprisingly, the result was the opposite of what was expected, as the incidence of late arrivals actually increased in the day-care center.

Experimental Design

Participants

The experiment was conducted in Georgia using the LIONESS software platform for interactive online experiments (Arechar et al., 2018). A total of 183 subjects participated, primarily from Tbilisi State University. We ensured that repeated participation was prevented by excluding duplicate IDs and IP addresses. Participants were not provided with information about the identities of their group members. Throughout all three treatments, the group members remained constant. Overall, 14 experiment sessions were carried out. The control treatment consisted of the standard PGG comprising 10 periods. The experiment had a duration of 10 to 20 minutes, and participants earned an average of 13.7 GEL (equivalent to 4.2 USD at that time). In the treatment involving a high centralized punishment probability mechanism and another treatment with a low centralized punishment probability mechanism, participants engaged in a two-stage PGG with 10 periods each. These sessions of the experiment lasted between 30 and 40 minutes, and participants earned an average of 20.7 GEL (equivalent to 6.3 USD) and 21.2 GEL (equivalent to 6.5 USD), respectively.

Method

A valuable tool for analyzing social dilemmas is the standard linear public goods game with a voluntary contribution mechanism (VCM) (Ledyard, 1995). As a control treatment, we conducted a standard linear PGG consisting of ten periods. To examine cooperative behavior in the presence of an external top-down incentive, we introduced a modified version of an exogenous centralized punishment mechanism with a probability, as elaborated in an experimental study by Stagnaro et al. (2017). In their study, different levels of probability for the exogenous centralized punishment mechanism were automatically introduced by a computer program using predetermined rules within the PGG. Each round of the PGG involved inspecting the contributions of players, and if a player did not fully contribute to the public goods, points were deducted.

In our case, with the punishment probability mechanism, we distinguished between inspection and penalty, assigning each of them their own probabilities. The reason for introducing separate inspection and penalty probabilities is to incorporate a more accurate understanding of the quality of the formal institution. The logic behind this approach is as follows: If someone is assigned to protect the provision of public goods but fails to discipline free riders simply because the opportunistic act was not observed, it is qualitatively different from the case where the opportunistic act is detected but not disciplined. The exogenous top-down centralized punishment mechanism serves as a demonstration of a formal institution. The higher the probability of punishing the free riders, the stronger and more trustworthy the formal institution and legal system are.

In our current study, we introduced a treatment with a weak formal institution (henceforth I&P9010). This treatment differs from treatment I&P9090 in the study by Mekvabishvili (2021a) only in terms of the level of probability of penalty.[2] Specifically, in stage 1 of treatment I&P9090, the probability

[2]The data set of our recent study Mekvabishvili (2021a) and our new experiment, including experimental instructions are available at Zenodo open data repository, Mekvabishvili, R. (2021b). Centralized Punishment in Public Good Experiments. Dataset, Zenodo, DOI: doi.org/10.5281/zenodo.5033369

Table 1. Design Information

Treatments	Stage 1 (periods 1-10)			Stage 2 (periods 11-20)			Number of Sessions	Number of Subjects
	Payoff Mechanism	Inspection Probability	Penalty Probability	Payoff Mechanism	Inspection Probability	Penalty Probability		
Control	VCM	0	0				3	57
I&P9090	CPM	90%	90%	VCM	0	0	6	64
I&P9010	CPM	90%	10%	VCM	0	0	5	65

of both inspection and penalty was 90%. In contrast, in treatment I&P9010, the probability of inspection remained high at 90%, but the probability of penalty was low, only 10%. Therefore, in our experimental setting, treatment I&P9010 allows us to examine the impact of a weak formal institution. In both treatments, in stage 2, CPM is removed, and subjects play a standard linear PGG consisting of ten periods. In both treatments, if the subjects were inspected and found to contribute less than the full endowment points, they were penalized. The penalty imposed was twice the number of points below the endowment point.

Thus, in the I&P9090 and I&P9010 treatments, we compare cooperation in the presence of strong and weak formal institutions and examine subsequent behavior after the CPM has been removed. We aim to measure the impact of the CPM on both choice behavior and intrinsic motivations. The behavior observed in both stages of the I&P9090 and I&P9010 treatments is then compared to that of subjects playing the standard linear PGG in the control treatment. Summary design information is presented in Table 1.

Payoff Mechanism

In the control treatment subject play a standard linear PGG for ten periods. In each period, subjects make simultaneous decisions regarding how much of their 20 endowment points to keep or invest into the public good. The payoff is determined by $\pi_i^1 = 20 - g_i + 0.375\sum_{j=1}^{n} g_i$, g_i is subject's contribution to public goods, and 0.375 is the marginal per-capita return of contributing to the public goods. The total payoff is the sum of the period payoffs over the ten periods. It is worth noting that full free-riding ($g_i = 0$) is a dominant strategy in the game. However, the aggregate payoff $\sum_{i=1}^{n} \pi_i^1$ is maximized if each group member fully cooperates ($g_i = 20$).

In the treatments with the CPM with probability, subjects play two-stage PGG with ten periods each. In stage 1, subjects in groups of four, play a standard linear PGG with the centralized punishment probability mechanism. In stage 1, the payoff is determined by $\pi_i^1 - 2 * (20 - g_i) * P(A|B) * P(B)$, where $\pi_i^1 = 20 - g_i + 0.375\sum_{j=1}^{n} g_i$, $P(A)$ is the probability that a penalty will be imposed, given the probability $P(B)$ that the contribution will be inspected, where $P(A)$ and $P(B)$ in treatment I&P9090 both equal to 0.9 and in treatment I&P9010 $P(A)$ is equal to 0.1 and $P(B)$ is equal to 0.9. In stage 2 of both treatments, the CPM is removed and the payoff is determined by $\pi_i^1 = 20 - g_i + 0.375\sum_{j=1}^{n} g_i$.

Information Conditions

In all three treatments, the composition of each group remained unchanged throughout the experiment. The players simultaneously made their contribution decisions, and once the decisions were made, they were informed about the contributions of their group members. However, in treatments I&P9090 and I&P9010, subjects were not informed about the inspection and penalty activities of their group members. They only knew about their own inspection and penalty activities in each period. To ensure that participants had consistent expectations regarding the length of the game, the total number of rounds was made known to all participants in all three treatments. Importantly, in treatments I&P9090 and I&P9010, the removal of the inspection and penalty mechanisms in stage 2 was not revealed to the participants in advance, but was introduced just before the start of stage 2. In order to ensure the quality of the data, we required participants to demonstrate comprehension of the game before playing the PGG. After reading the instructions, participants were unable to proceed to the game until they answered all control questions correctly (with an unlimited number of attempts allowed).

Results

Stage 1 of the treatments with the CPM

The presence of the CPM leads to a significant increase in average contribution levels in both treatments. In stage 1 of treatment I&P9090, the average contribution rate is 92% (18.5 points, standard deviation 0.66) of the endowment, while in treatment I&P9010, it is 82% (16.4 points, standard deviation 1.1). In both treatments, the mean contribution starts at a relatively high level. However, in treatment I&P9090, the mean contribution steadily converges towards socially beneficial equilibria, while in treatment I&P9010, it starts to diverge from the path to socially beneficial equilibria after period four and declines steadily. The difference in mean contributions between treatments I&P9010 and I&P9090 over all ten periods of stage 1 is significant (Mann-Whitney test, p = 0.0003). This evidence suggests that different levels of punishment probability in stage 1 have a differential impact on decisions.

Nevertheless, despite the low probability of punishment, the I&P9010 treatment still achieves a relatively high level of cooperation.

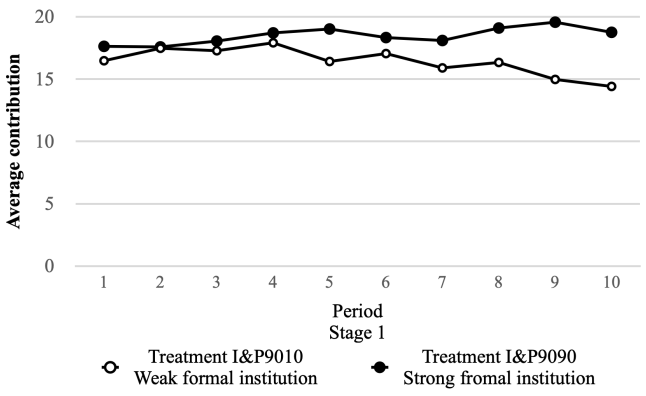

Figure 1. Cooperation under high and low CPM

Result 1: *In stage 1 of the I&P9090 treatment, average contributions converge to socially beneficial equilibria over time. However, in the I&P9010 treatment, average contributions remain substantially high but exhibit divergence from socially beneficial equilibria starting from period four.*

To provide formal statistical evidence for Result 1, we conducted a regression analysis of cooperative behavior under the CPM. Table 2 presents the model and the ordinary least-squares (OLS) regressions separately for the I&P9010 and the I&P9090 treatment. In Table 3, we present estimates for the effect of the CPM on contributions at the subject level. A simple model that captures time effects is used, where contributions are estimated as a function of the "Period" (i.e., the period number).

The variable "Inspected" is a dummy variable that takes the value 1 if a subject was inspected and 0 otherwise. The variable "Penalty" indicates the penalty points incurred by the subject. Additionally, the variables "Average contribution" and "Average payoff" serve as group control variables, representing the group's average contributions and average payoffs, respectively.

The regression results suggest that in the I&P9090 treatment, subjects contribute less when they are punished. In the I&P9010 treatment, the coefficient for the variable "Period" is negative, indicating a decay in cooperation over time. Since both treatments had a high probability of inspection, but a very low probability of penalty in the I&P9010 treatment, the variable "Inspection" had a negative impact on contributions, suggesting that when inspection is not followed by punishment, it has a weaker disciplining effect.

In both treatments, the coefficient for "Average contribution" is positive and highly significant. This indicates that as the average contribution of other group members increases, individuals tend to contribute more. It is worth noting that in the I&P9010 treatment, the variable "Average payoff" negatively affects contribution levels, but does not have a significant impact, while in the I&P9090 treatment, it has a significant impact. This result suggests that individual decisions in the I&P9010 treatment were not significantly influenced by the financial outcomes of other group members, likely due to the virtual absence of punishment, whereas in the I&P9090 treatment, the observed losses from punishment did have an impact.

Table 2. Results from linear regressions on contribution decisions under CPM

Dependent variable: Contributions			
Independent variables		I&P9090	I&9010
Constant		15.1540	0.1708
		(1.0916)	(3.6228)
Average contribution		0.5798*	1.0012*
		(0.1007)	(0.1761)
Average payoff		-0.2360*	-0.0005
		(0.0432)	(0.1180)
Period		0.0106	-0.0013
		(0.0270)	(0.0921)
Inspected		0.1497	-0.1778
		(0.2403)	(0.7542)
Penalty		-0.4910*	na†
		(0.0093)	
	N	640	660
	Adjusted R^2	0.834	0.068
	F	644.27*	5.88

Note: Standard errors are in parentheses. * denotes significance at the 5-percent level.† since the punishment probability was low (10%) in I&P9010 treatment, the variable "penalty" was zero in our data set, as in all sessions of the treatment no single free-rider happened to be penalized.

We compared the mean contributions in the control treatment with those in the I&P9090 and I&P9010 treatments to evaluate the differences. Figure 2 illustrates the differences between the I&P9090 and I&P9010 treatments compared to the control treatment.

In the control treatment, the mean contributions exhibit a surprisingly consistent pattern, with voluntary contributions remaining relatively high throughout all rounds, well above 50% of the endowment. The average contribution rates amount to 70% (14.1 points, standard deviation 0.97) of the endowment. However, in the last round, which is typical for this type of public goods experiment, there is a pronounced endgame effect with a sharp drop in contributions (Wilcoxon signed-rank test, two-sided, differences between round 1 and round 10, $p = 0.000$). In the last period, 22% of the subjects contributed zero points.

Despite the relatively high contribution levels in the con-

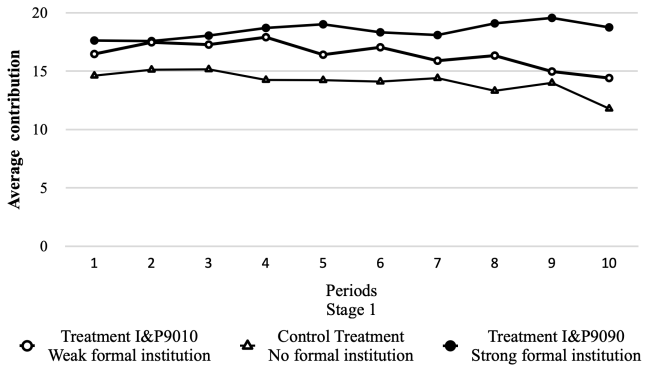

Figure 2. Cooperation with and without CPM

trol treatment, the average contributions in stage 1 of both the I&P9090 and I&P9010 treatments are significantly higher (Mann-Whitney test, two-sided, I&P9010 treatment p = 0.000, I&P9090 treatment p = 0.000). This suggests that regardless of whether the centralized punishment probability mechanism is high or low, contributions are significantly higher compared to the no-punishment case.

Result 2: *Although the mean contributions in the control treatment are relatively high and stable, the introduction of both low and high CPM leads to significantly higher contribution levels on average. The CPM proves effective in maintaining high cooperation levels in both the I&P9090 and I&P9010 treatments. However, average contributions are significantly higher in the I&P9090 treatment.*

Stage 2 of the treatments with CPM

We observed a significant decrease in contributions and, consequently, inefficiency in the second stages. Table 3 presents the mean, standard deviation, and median contributions at the group level in the I&P9090 and I&P9010 treatments. In the I&P9090 treatment, average contributions in stage 1 are 53% lower compared to stage 2, while in the I&P9010 treatment, they are 44% lower. Once CPM is removed in stage 2, the mean contributions in period 11 start at low levels relative to stage 1 in both treatments. They then converge to Nash equilibria of zero contribution (the differences between stage 1 and stage 2 in both the I&P9010 and I&P9090 treatments were significant at $p < 0.05$ (two-sided) according to a Wilcoxon matched-pairs test).

While the average contributions in the I&P9010 and I&P9090 treatments stabilize around 16 and 18, respectively, in stage 1, there is an immediate and significant drop in contributions in period 11 (Wilcoxon signed rank test, two-sided, p = 0.000). Moreover, we found that contributions in period 11 differed significantly across individuals in both the I&P9010 and I&P9090 treatments (Chi Squared test, p = 0.000). This indicates that the removal of the punishment mechanism triggers forces that strengthen the equilibrium of complete free-riding.

Table 3. Average contributions

	Treatments			
	I&P9010		***I&P909010***	
	Stage 1 CPM	***Stage 2*** VCM	***Stage 1*** CPM	***Stage 2*** VCM
Mean	16.5	9.3	18.4	8.6
Standard deviation	2.4	3.3	1.7	3.5
Median	17.3	7.9	20.0	6.7
N (independent groups)	18	18	18	18

It is worth noting that in the I&P9010 treatment, in stage 1, 71% of subjects who contributed 50% or more of their endowment in at least 8 out of 10 periods maintained the same high contribution level in period 11. Similarly, in the I&P9090 treatment, in stage 1, 88% of subjects who contributed 50% or more of their endowment in at least 8 out of 10 periods also maintained the same high contribution level in period 11. In stage 2 of the I&P9090 treatment, the mean contributions decline and reach 7.4 in period 20. A similar pattern of cooperation is observed in stage 2 of the I&P9010 treatment, where the mean contributions reach 7.2 in the last period. In stage 2, there is no significant difference in mean contributions between the I&P9090 and I&P9010 treatments (Mann-Whitney test, two-sided, p = 0.44). This evidence indicates that different probability levels of the CPM in stage 1 do not have a differential impact on decisions in stage 2.

Result 3: *When the CPM is removed in stage 2 of both the I&P9090 and I&P9010 treatments, average contributions decrease significantly and converge towards full free-riding. However, in the presence of the CPM, the cooperation levels remain high.*

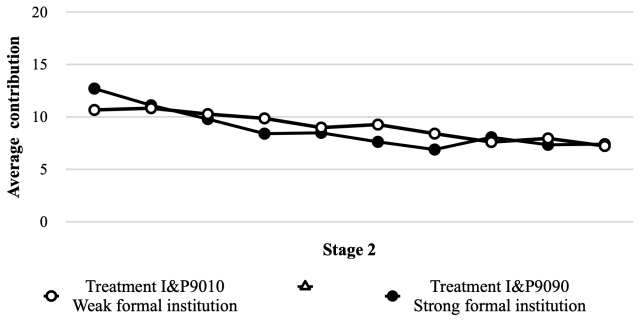

Figure 3. Contributions after removal of CPM

Result 4: *In treatments I&P9010 and I&P9090, the levels of cooperation in stage 2 are similar to each other and significantly lower compared to stage 1. Furthermore, they both converge towards a state of free-riding over time.*

We compared the mean contributions of the control treatment to those in stage 2 of the I&P9090 and I&P9010 treatments. Although the I&P9090 and I&P9010 treatments consisted of two stages while the control treatment was a single-stage standard PGG, we placed the results of the control treatment (represented by a dotted line) into stage 2 of the I&P9090 and I&P9010 treatments for better comparison. Figure 4 illustrates this comparison.

In both treatments with the CPM, the mean contributions are significantly lower than in the control treatment (Mann-Whitney test, two-sided, p = 0.000). This result suggests that the difference in cooperation levels can be attributed to the experience of the CPM in stage 1 of the I&P9090 and I&P9010 treatments. We observe a significantly lower level of cooperation in stage 2 of the treatments with both low and high probabilities of the CPM compared to the cooperation level in the control treatment with no punishment mechanism.

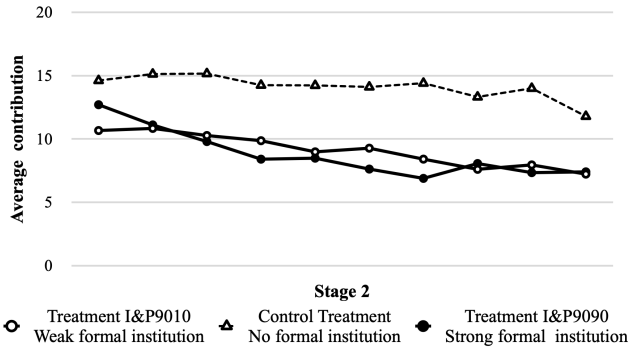

Figure 4. Cooperation with and without experience of the CPM

Result 5: *In treatments I&P9010 and I&P9090, the levels of cooperation in stage 2 are significantly lower compared to the control treatment, eventually converging to a state of free-riding over time.*

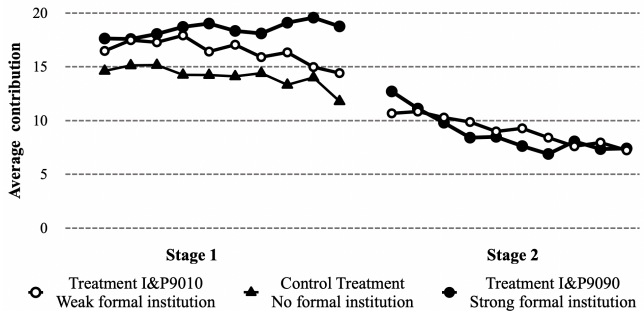

Figure 5. Cooperation dynamics under low and high CPM presence and subsequent removal

Figure 5 summarizes the results of all three treatments, allowing us to track the dynamics of cooperative behavior in the presence and absence of the CPM and compare it to the cooperation level of the control treatment with no punishment mechanism.

Welfare effects

We closely examine the penalty cases and their magnitude in stage 1 of the I&P9090 and I&P9010 treatments. In the I&P9010 treatment, where the probability of CPM was low, no penalty cases were recorded. As a result, we observed a relatively higher frequency of moderate-sized risky choices. In the I&P9090 treatment, despite the high probability of CPM, risky decisions were still made in each period, albeit with a decreasing trend. Figure 6 illustrates the progression of penalty cases in the I&P9090 treatment. Since the penalty amounted to twice the points that players kept for themselves, players took small risks and kept small amounts, averaging 1.2 points (6% of the total endowment points).

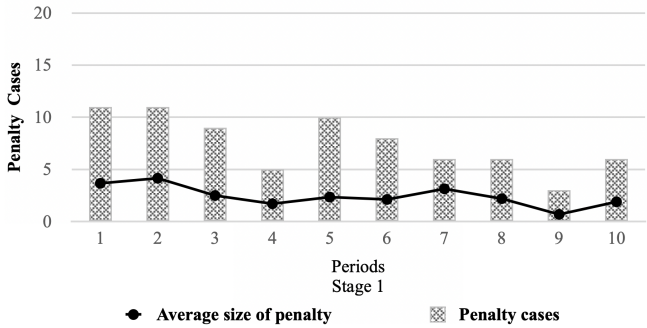

Figure 6. Average penalty size and cases under the strong formal institution

Result 6: *In the treatment with a high probability of CPM, the prevalence of full contributions is the highest. However, selfish decisions persist even in the presence of high CPM probabilities, but they tend to be smaller in size and exhibit a declining trend on average. Conversely, in the treatment with a lower probability of CPM, we observe a greater frequency of risky decisions.*

Next, we examine whether the CPM also improves net earnings and whether there are differences between the weak formal and strong formal institution treatments. Additionally, we are interested in the subsequent welfare development in stage 2 after the removal of the CPM in both the I&P9090 and I&P9010 treatments. Figure 7 displays the percentage-based development of average per-period individual net earnings over time.

Welfare is measured by the average individual net earnings per period, which is the earnings after deducting the received penalty. The group welfare-maximizing level of contribution is the full contribution of 20 points by all four members of the group, resulting in each group member earning 30 points in all three treatments.

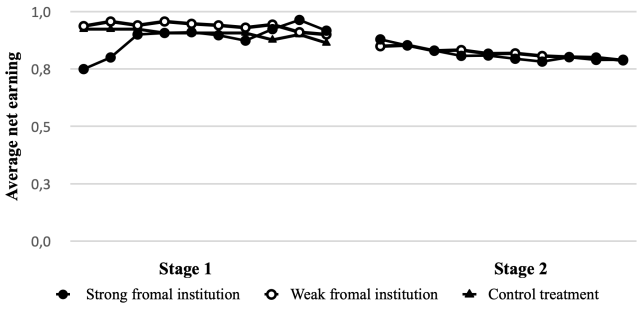

Figure 7. Average per–period individual net earnings

Result 7: *Welfare, measured by the average per-period net earnings, is higher in the I&P9010 treatment compared to the I&P9090 and control treatments. However, after the removal of the CPM in stage 2, the welfare decreases relative to stage 1.*

In the left panel of Figure 7, the average per-period individual net earnings of all three treatments are shown to be quite high. It can be observed that the disadvantage of the I&P9090 treatment compared to the I&P9010 treatment diminishes over time, as the average per-period net earnings in both treatments converge in the last 7 periods of stage 1. However, the average per-period net earnings in stage 1 of the I&P9090 treatment are significantly lower than in the I&P9010 treatment (Mann-Whitney test, two-sided, p=0.0126). While there is no significant difference between the control treatment and the I&P9090 treatment, the average per-period net earnings in stage 1 of the I&P9010 treatment are significantly higher than in the control treatment (Mann-Whitney test, two-sided, p=0.002). Thus, the relatively high penalty cases in the early periods of the game are primarily responsible for the decreased efficiency in the I&P9090 treatment. Interestingly, in the I&P9010 treatment, where the penalty probability is relatively low, high levels of welfare are achieved, which is encouraging.

In the right panel of Figure 7, there is no significant difference in average per-period net earnings between the I&P9090 and I&P9010 treatments in stage 2 (Mann-Whitney test, two-sided, p = 0.4059). In the I&P9090 treatment, there was no significant difference in average net earnings between stages (not significant at $p<0.05$, two-sided, according to a Wilcoxon matched-pairs test). However, in the I&P9090 treatment, net earnings appeared to be significantly higher in stage 1 than in stage 2 (significant at $p<0.05$, two-sided, according to a Wilcoxon matched-pairs test).

Discussion

Our evidence suggests that strong top-down institutional incentives to cooperate have an effective disciplining impact on free-riding, and a high level of cooperation is maintained. Interestingly, even in the case of a weak formal institution remains at a relatively high level. This could be attributed to the setup of the centralized punishment mechanism, where inspection and punishment are detached. It seems that even regular inspection alone serves as a deterrent for free-riding behavior to some extent. Although this is an encouraging observation, our results indicate that over time, if free-riding behavior is only revealed and rarely penalized, cooperation starts to decay. As a result, when punishment becomes less credible, free riders are more incentivized to cheat, while cooperative individuals become more cautious, leading to decreased contributions to avoid being exploited by free riders.

Turning to the welfare effects, net earnings are significantly lower in the case of a strong formal institution compared to a weak formal institution. However, net earnings quickly converge to similarly high levels over time. The large differences in earnings at the initial periods can be attributed to the construction of the centralized punishment mechanism. In the case of a strong formal institution, a number of subjects initially attempt to free ride, but once they are penalized, they increase their contribution levels. We observe similar behavior in the case of a weak formal institution, but since free riders are only inspected and rarely penalized, no welfare losses are incurred. Hence, one policy implication could be that a warning system can be effective in limiting opportunistic behavior at the initial stage. However, if it is not supported and followed by a credible punishment mechanism, it may prove to be largely inefficient.

Our experimental results demonstrate that exposure to both strong and weak formal institutions, which provide top-down motivation for cooperation, does not lead to increased prosocial behavior after their removal. One possible explanation for the absence of evidence regarding the spillover effect into subsequent stages without punishment could be the single domain nature of the experiment. Another possible reason could be the insufficient exposure of the subjects to top-down incentives for cooperation. Future research should therefore aim to explore how varying lengths of exposure to such incentives can result in different effect sizes.

On the other hand, shifting from the spillover effect to the crowding out effect, our findings align with a large body of evidence on "crowding out" effects, where internal motivations to achieve a certain goal can be replaced by external incentives. We observe that both harsh and softer punishment mechanisms (i.e., those with higher or lower probabilities of penalizing non-contribution) lead to a crowding out effect. In this regard, a previous experimental study by Frohlich and Oppenheimer (2003) supports our findings. They argue that exogenous incentives remove the need for individuals to reason and enforce cooperation themselves, stating: "They don't have to flex their ethical muscles" (Frohlich and Oppenheimer, 2003, p. 290). The intuition behind this is straightforward. If individuals develop cooperation under an external enforcement system that is later lifted, they become extremely cautious about being exploited by others. They would prefer to cooperate if they knew their group members were also willing to cooperate. However,

the process of mutual learning about group member types and preferences is hindered under the shadow of an external incentive. As a result, the building of interpersonal trust is discouraged, as each group member is more likely to attribute cooperation to the external institutional incentive rather than to the benign intentions and beliefs of their fellow members. This result also extends to the cases where public goods are provided by public-private partnerships (see Martimort and Pouyet, 2008 from a theoretical point of view and Attanasi et al., 2020, for recent experimental evidence).

In general, cooperative norms contribute to a society's "social capital" and can enhance allocative efficiency by reducing monitoring and contract enforcement costs. Norms of civic cooperation are social norms that restrain individuals' narrow self-interest and facilitate the provision of public goods. Examples include norms against littering, abusing the welfare state, or evading fares on public transport. The absence of a spillover effect and the presence of a crowding out effect on cooperation suggest that the strength of institutions does not influence prosocial behavior through a change in perceived social norms.

If exposure to strong or weak institutions were to influence prosociality by altering people's explicit understanding of appropriate behavior (i.e., their perception of social norms), it would also result in changes in cooperative behavior. It is evident that policies promoting a more cooperative environment are cost-effective. However, policymakers should approach policy design with caution. Furthermore, when the legal or regulatory framework - the "institutional environment" - lacks credibility, individuals are more likely to behave opportunistically and make less efficient adjustments to government policies.

Conclusion

Our results suggest that the application of exogenous centralized punishment as a policy tool in social dilemmas can be a two-edged sword. Regardless of whether the formal institution is strong or weak, the removal of this external incentive can undermine the level of cooperation to a similar extent. In other words, both weak and strong formal institutions imposed externally lead to a crowding out effect on individuals' intrinsic motivation to cooperate. Therefore, based on the current experimental evidence, it is challenging to achieve both a socially beneficial outcome and intrinsically motivated cooperation simultaneously through exogenous top-down centralized punishment in the single context of the public goods game. Furthermore, our findings indicate that a higher level of cooperation can be achieved among subjects who do not experience external top-down motivation compared to those who have had such an experience. While our study contributes to the experimental research on the role and impact of institutions on cooperative behavior in social dilemmas, it also highlights the need for further research and emphasizes the importance of a comprehensive and cautious approach to policy design.

Acknowledgments

We would like to thank Prof. Antonio Cabrales and Dr. Benedikt Herrmann for their comments. This work was supported by Shota Rustaveli National Science Foundation (SRNSF), grant PHDF-19-321, "Prosocial Behavior in Economics and Influence of State Institutions: Case of Georgia".

References

Arechar, A. A., Gächter, S. & Molleman, L. (2018). Conducting interactive experiments online. *Experimental Economics, 21*(1), 99-131.

Attanasi, G., Boun My, K., Buso, M., & Stenger, A. (2020). Private investment with social benefits under uncertainty: The dark side of public financing. *Journal of Public Economic Theory, 22*(3), 769-820.

Baldassarri, D., & Grossman, G. (2011). Centralized sanctioning and legitimate authority promote cooperation in humans. *Proceedings of the National Academy of Sciences, 108*(27), 11023-11027.

Bénabou, R., & Tirole, J. (2006). Incentives and prosocial behavior. *American Economic Review, 96*(5), 1652-1678.

Bowles, S. & Polania-Reyes, S. (2012). Economic incentives and social preferences: substitutes or complements? *Journal of Economic Literature, 50*(2), 368-425.

Engl, F., Riedl, A., & Weber, R. (2021). Spillover effects of institutions on cooperative behavior, preferences, and beliefs. *American Economic Journal: Microeconomics, 13*(4), 261-99.

Fehr, E. & Gachter, S. (2000). "Cooperation and Punishment in Public Goods Experiments." *American Economic Review, 90*(4), 980-994.

Frey, B. (2017). Policy consequences of pay-for-performance and crowding-out. *Journal of Behavioral Economics for Policy, 1*(1), 55-59.

Frey, B. S. & Jegen, R. (2001). Motivation crowding theory. *Journal of Economic Surveys, 15*(5), 589-611.

Frohlich, N. & Oppenheimer, J. (2003). Optimal policies and socially oriented behavior: Some problematic effects of an incentive compatible device. *Public Choice, 117*(3), 273-293.

Funk, P. (2007). Is there an expressive function of law? An empirical analysis of voting laws with symbolic fines. *American Law and Economics Review, 9*(1), 135-159.

Galbiati, R. & Vertova, P. (2014). How laws affect behavior: Obligations, incentives, and cooperative behavior. *International Review of Law and Economics, 38*, 48-57.

Gneezy, U., Meier, S. & Rey-Biel, P. (2011). When and why incentives (don't) work to modify behavior. *Journal of Economic Perspectives, 25*(4), 191-210.

Gneezy, U. & Rustichini, A. (2000). Pay enough or don't pay at all. *The Quarterly Journal of Economics, 115*(3), 791-810.

Herrmann, B., Thoni, C., & Gachter, S. (2008). Antisocial punishment across societies. *Science, 319*(5868), 1362-1367.

Ledyard, J. O. (1995). "Public goods: A survey of experimental research." In *The Handbook of Experimental Economics*, 111–194 (Ed.: J. Kagel & A. E. Roth). Princeton University Press.

Martimort, D., & Pouyet, J. (2008). To build or not to build: Normative and positive theories of public–private partnerships. *International Journal of Industrial Organization, 26*(2), 393-411.

Mekvabishvili, R. (2021a). Can Formal Institutions Lead to the Spillover Effect of Cooperation? *Theoretical Economics Letters, 11*, 186-193.

Mekvabishvili, R. (2021b). Centralized Punishment in Public Good Experiments. Dataset, *Zenodo*, DOI: doi.org/10.5281/zenodo.5033369

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.

O'Gorman, R., Henrich, J., & Van Vugt, M. (2009). Constraining free riding in public goods games: designated solitary punishers can sustain human cooperation. *Proceedings of the Royal Society B: Biological Sciences, 276*(1655), 323-329.

Peysakhovich, A., & Rand, D. G. (2016). Habits of virtue: Creating norms of cooperation and defection in the laboratory. *Management Science, 62*(3), 631-647.

Putterman, L., Tyran, J. R., & Kamei, K. (2011). Public goods and voting on formal sanction schemes. *Journal of Public Economics, 95*(9-10), 1213-1222.

Stagnaro, M., N., Arechar, A., A. & Rand, D., G.; (2017). From Good Institutions to Generous Citizens: Top-Down Incentives to Cooperate Promote Subsequent Prosociality But Not Norm Enforcement. *Cognition, 167*, 212-254.

Tyran, J. R., & Feld, L. P. (2006). Achieving compliance when legal sanctions are non-deterrent. *Scandinavian Journal of Economics, 108*(1), 135-156.