Economics Letters 268 (2026) 113227

Contents lists available at ScienceDirect

Economics Letters

journal homepage: www.elsevier.com/locate/ecolet

Weaponized reputation: Experimental evidence on the misuse of coordinated punishments ☆

Haihan Yu

School of Economics, Zhejiang University of Finance and Economics, China

ARTICLE INFO

JEL classification:
C73
C92
L14

Keywords:
Coordinated punishments
Relational contracts
Repeated games

ABSTRACT

This paper experimentally studies how coordinated punishment can be misused. In a repeated trust game with public reputation, we compare a silent commitment treatment, where Agents privately choose a review threshold (the minimum payment required for a positive review), with a public commitment treatment, where the review threshold is observable before Principals choose payments. Public commitment significantly reduces cooperation: effort and payments fall relative to silent commitment. Public commitment also turns posted review thresholds into more enforceable demands: Principals are substantially more likely to meet visible thresholds. Successful zero-effort extraction, in which an Agent exerts no effort but sets a positive review threshold met by the Principal's payment, is significantly more frequent under public commitment. Coordinated punishments can therefore discipline misconduct, but may also create tools for extortion.

1. Introduction

Repeated interaction can support efficient outcomes that are unavailable in one-shot play, but participants must coordinate on the same equilibrium. In the repeated prisoner's dilemma, this usually means coordinating on cooperative rather than noncooperative play. Asymmetric relationships add another issue because equilibrium selection also determines how the gains from cooperation are divided. Making a binding commitment observable can help coordination by revealing one player's contingent response before the other acts. Yet the player making the commitment also chooses the condition that triggers that response. If punishment is tied to a payment demand rather than to productive behavior, the other player may satisfy the demand even when little effort is provided. Commitment can therefore coordinate payment while weakening the link between effort and reward.

Reputation systems make this possibility concrete. Future trading partners can punish past defection, allowing reputation to sustain anonymous trade (Greif et al., 1994). Because a current review helps determine whether future partners impose that punishment, it also gives the reviewer leverage over the current partner. Retaliatory reviews on Airbnb (Conti, 2019) and eBay (Klein et al., 2016), along with grading incentives created by student evaluations (Langbein, 2008; Krautmann and Sander, 1999), illustrate how suppliers of reputational information can use this leverage. In each case, the same information that coordinates later responses also gives its supplier influence over current terms.

Barron and Guo (2021) formalize the leverage created by this structure. The long-lived Principal relies on reviews to coordinate punishment by future short-lived Agents. If the current Agent's binding review rule is observed before payment, the Agent can threaten a bad review unless the payment meets a chosen threshold. Because this threshold need not depend on effort, the review rule can enforce payment even after the Agent shirks. The model delivers an extreme prediction: this possibility unravels the relational contract, producing zero effort and zero payment.

We test this prediction in a repeated trust game with public reputation. The design builds on the repeated trust-game experiment of Brown et al. (2004). In each round, an Agent chooses effort and a payment-contingent review rule. The Principal then chooses a payment, and the resulting review enters the Principal's public record. The review rule is binding in both treatments. In the Silent Commitment Treatment, the Principal does not observe it before paying. In the Public Commitment Treatment, the Principal does. The comparison therefore holds commitment fixed and isolates whether the Agent's payment condition is visible when the Principal chooses payment. The resulting theoretical contrast is sharp: hidden commitment can sustain an efficient relational contract, whereas observable commitment predicts complete collapse. This setting allows us to test whether making commitment observable supports productive exchange or turns coordinated punishment into leverage for a payment demand unrelated to effort.

☆ I am grateful to the editor, Professor Joseph Harrington, and the referee for their constructive comments and suggestions, which substantially improved the paper. I thank Feng Chen for the excellent research assistance. This work was supported by the National Natural Science Foundation of China (72203199).

E-mail address: haihanyu@zufe.edu.cn.

https://doi.org/10.1016/j.econlet.2026.113227
Received 31 July 2026; Received in revised form 1 September 2026; Accepted 3 September 2026
Available online 10 September 2026

Table 1
Effort cost schedule in the stage game.

Effort e_t	0	1	2	3	4	5	6	7	8	9	10	11	12
Cost $c(e_t)$ (ECU)	0	1	3	6	10	15	21	28	36	45	55	66	78

2. Experimental design

We design a laboratory experiment using a repeated trust game with a reputation system. Participants are randomly assigned to a role of Principal or Agent, fixed throughout the experiment.

The stage game

Each round proceeds in three steps. Both players receive a 40 Experimental Currency Unit (ECU) endowment per round. Every realized main-task round is paid.

Step 1: The agent chooses effort. The Agent selects effort $e_t \in \{0, 1, \dots, 12\}$. Each unit generates 8 ECU for the Principal, $R(e_t) = 8e_t$. The convex cost function $c(e_t)$ is given by Table 1.

Step 2: The principal chooses a transfer. After observing the Agent's effort e_t, the Principal chooses a monetary transfer $s_t \in \{0, 1, \dots, 100\}$ ECU to send to the Agent.

Step 3: The agent leaves a public review. The Agent sends a public message $m_t \in \{\text{Good}, \text{Bad}\}$ about the Principal. The message is recorded on a public board: every later Agent matched with this Principal in the current supergame observes the Principal's full message history.

The resulting single-round payoffs are:

Agent's payoff: $\pi_t^A = 40 - c(e_t) + s_t$

Principal's payoff: $\pi_t^P = 40 + 8e_t - s_t$

Strategy method

We use the Strategy Method to elicit the Agent's review profile. In both treatments, the Agent simultaneously chooses effort e_t and a review profile μ_t: a rule specifying a review threshold $\bar{s}_t$ — the minimum payment required to earn a Good review — that returns Good if $s_t \geq \bar{s}_t$ and Bad otherwise. The message m_t is then automatically determined by $\mu_t(s_t)$, making the Agent's review threshold a binding commitment device.

Repeated interaction and reputation building

We implement the infinite horizon via random termination. Each session has five supergames; within a supergame, each round continues with probability $p = 0.80$ and terminates with probability 0.20. If the five supergames are not completed within 40 stage-game rounds in total, the main task ends after round 40. When a supergame ends, all participants advance together, and reputation records and matching history reset. Within a supergame, every Agent observes the matched Principal's full message history, enabling coordinated punishment of Principals with poor reputations.

Treatments

The two treatments share stage game, strategy method, and repeated-interaction structure; they differ only in the Principal's information set before choosing s_t:

Treatment 1 — Silent Commitment Treatment (SCT): The Principal observes the Agent's effort e_t but does *not* observe the review profile μ_t. Because the review threshold is hidden, it cannot function as a pre-payment threat.

Treatment 2 — Public Commitment Treatment (PCT): The Principal observes both the Agent's effort e_t and the review profile μ_t. The Agent's review threshold is therefore a publicly observable and action-contingent threat.

The SCT is the benchmark for standard coordinated punishment; PCT introduces the extortion vulnerability that is the focus of our predictions.

The Agent's stated review threshold is binding under the Strategy Method in both treatments. What changes is whether the Principal sees it before choosing the transfer. In SCT, the Principal does not see the threshold and must form beliefs about the Agent's review rule. This is the timing in the standard coordinated-punishment model, where the review is communicated only after payment. A threshold that is hidden at the payment stage cannot be used to demand a transfer. In PCT, the Principal sees the threshold first, allowing the Agent to tie the review to a stated payment demand.

Risk elicitation and payment

After the main task, participants complete the Bomb Risk Elicitation Task (Crosetto and Filippin, 2013). Participants choose how many of 100 boxes to collect; each box pays 1 ECU unless the bomb is among the boxes collected, in which case the task pays zero. Final payment consists of the ECU payoffs from every realized main-task round and risk-task earnings.[1] Experimental earnings are converted at a rate of 16 ECU to 1 RMB.

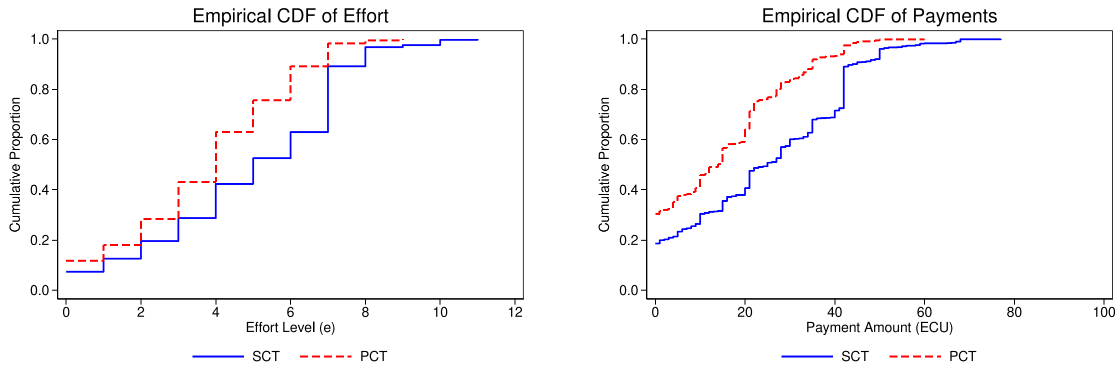
3. Theoretical framework

Following Barron and Guo (2021), we formalize the interaction. Let P denote the long-lived Principal and A_t the short-run Agent in period $t \in \{0, 1, \dots\}$. The continuation probability (discount factor) is $\delta = 0.80$. In each period, the Agent chooses an effort level $e_t \in \{0, 1, \dots, 12\}$, which generates a benefit $R(e_t) = 8e_t$ for the Principal. The cost function $c(e_t)$ is summarized in the effort-cost schedule shown in Table 1. The transfer $s_t \in \{0, 1, \dots, 100\}$ and $m_t \in \{Good, Bad\}$. $c(e_t)$ is strictly increasing and convex (marginal differences $\Delta c(e_t)$ are positive and increasing). Because effort is discrete, total surplus $W(e) = 8e - c(e)$ is maximized at both $e = 7$ and $e = 8$: $W(7) = 56 - 28 = 28$ and $W(8) = 64 - 36 = 28$. Thus, the first-best effort set is $E^{FB} = \{7, 8\}$. Stage payoffs are $\pi_t^P = 8e_t - s_t$ and $\pi_t^A = s_t - c(e_t)$. The Principal's normalized continuation payoff is $\Pi_t = (1-\delta)\sum_{\tau=t}^{\infty} \delta^{\tau-t}(8e_\tau - s_\tau)$. Let $\overline{\Pi}$ and $\underline{\Pi}$ be the maximum and minimum continuation payoffs. The Agent's maximum *Leverage* is $L = \frac{\delta}{1-\delta}(\overline{\Pi} - \underline{\Pi}) = 4(\overline{\Pi} - \underline{\Pi})$. We solve for the Perfect Bayesian Equilibrium (PBE) of each communication regime.

Treatment 1: Silent commitment (Sustained cooperation)

In the SCT, A_t simultaneously chooses e_t and a review profile $\mu_t(s_t)$, but P observes only e_t before choosing s_t. The hidden commitment

[1] Before collecting the sample analyzed here, we ran four sessions with 126 participants using the same treatments and procedures except for the main-task payment rule. One supergame was selected at random, and each participant's main-task earnings equaled the average per-round payoff in that supergame. As the referee pointed out, this averaging rule makes effective discounting vary across rounds and therefore does not implement the stationary repeated game with $\delta = 0.80$ analyzed here. We consequently exclude those observations from the present analysis. The qualitative treatment pattern is the same: effort and payments are lower in PCT than in SCT, although the average levels of both outcomes are lower in the earlier sample. For a detailed discussion of payment rules in repeated-game experiments, see Fréchette et al. (2025). We thank the referee for bringing this issue to our attention.



(a) Effort (b) Payment

Fig. 1. Empirical cumulative distributions of effort and payments by treatment. Each curve gives the proportion of observations weakly below the value on the horizontal axis and pools all rounds for the relevant role within each treatment. SCT is represented by the solid blue line and PCT by the dashed red line.

cannot threaten extraction and is strategically equivalent to choosing m_t after observing s_t: communication is cheap talk. Cooperation can therefore be sustained by a grim-trigger profile in which P pays s^* if $e_t \geq e^*$ and 0 otherwise; A_t exerts e^* and sets $\mu_t(s_t) = Good$ if $s_t \geq s^*$, Bad otherwise. Any Bad message triggers reversion to the static Nash $e_t = 0, s_t = 0$, yielding $\underline{\Pi} = 0$. A_t exerts e^* if and only if the transfer weakly exceeds the cost:

$$s^* - c(e^*) \geq 0 \implies s^* \geq c(e^*)$$

The Principal pays s^* if the immediate cost is offset by the cooperative continuation payoff $\overline{\Pi} = 8e^* - s^*$. The enforcement constraint is:

$$(1-\delta)(-s^*) + \delta\overline{\Pi} \geq 0 \implies s^* \leq \frac{\delta}{1-\delta}\overline{\Pi} = 4(8e^* - s^*)$$

$$5s^* \leq 32e^*$$

We focus on principal-optimal efficient equilibria, which provide the upper bound on cooperation against which the extortionate collapse contrasts. The Principal minimizes s^* subject to the Agent's participation constraint, so $s^* = c(e^*)$. Since $E^{FB} = \{7, 8\}$, this construction yields two principal-optimal efficient contracts:

$$(e^*, s^*) \in \{(7, 28), (8, 36)\}.$$

Both satisfy the enforcement constraint strictly:

$$5 \times 28 = 140 < 224 = 32 \times 7, \qquad 5 \times 36 = 180 < 256 = 32 \times 8.$$

Thus, the principal-optimal efficient equilibrium is set-valued. The $(7, 28)$ contract has greater enforcement slack, but both contracts deliver the same on-path payoffs, $(\pi^P, \pi^A) = (28, 0)$, and the same maximum per-round surplus of 28 ECU. This common surplus serves as the welfare benchmark against which empirical efficiency losses are measured.

Treatment 2: Public commitment (Collapse of cooperation)

In the PCT, the Principal observes $\mu_t(s_t)$. The Agent publicly commits to Bad for transfers below a review threshold $\hat{s}$ and $Good$ otherwise. Faced with this visible threat, the Principal pays $\hat{s}$ if and only if $\hat{s} \leq L$. Because $\hat{s}$ depends only on the Principal's continuation payoff and is disjoint from e_t, the Agent solves:

$$\max_{e_t \in \{0,1,\ldots,12\}} \hat{s} - c(e_t) \quad \text{subject to} \quad \hat{s} \leq L$$

Since $c(e_t)$ is strictly increasing, the Agent sets $e_t = 0$ for any achievable $\hat{s}$. With zero effort on the equilibrium path, $\overline{\Pi} = 0$ and $L = 0$. The unique equilibrium outcome is a complete collapse: $e^* = 0$, $s^* = 0$.

Hypotheses

Based on the theoretical predictions, we test two hypotheses:

Hypothesis 1 (*Cooperation Collapse*)**.** Average effort and payment in Treatment 2 are lower than in Treatment 1; observable action-contingent threats break the relational contract.

Hypothesis 2 (*Emergence of Extortion*)**.** Under public commitment, the Agent's review threshold becomes a visible, more enforceable payment condition rather than a private target. We therefore expect Principals' payments in Treatment 2 to more often meet posted thresholds, and some Agents to extract payments even after choosing zero effort.

4. Results

We conducted the experiment at Zhejiang University of Finance and Economics with 134 student participants, none of whom had previously participated in a similar experiment. Participants were randomly assigned to one of six sessions: 72 in SCT (three sessions) and 62 in PCT (three sessions). Each participant received a printed copy of the instructions, which were read aloud. Participants could proceed to the main task only after answering every question in a comprehension quiz correctly.[2] The experiment was programmed in oTree (Chen et al., 2016). The average age of participants was 21.93 years; sessions lasted about 75 min; average earnings were 77.45 RMB.[3]

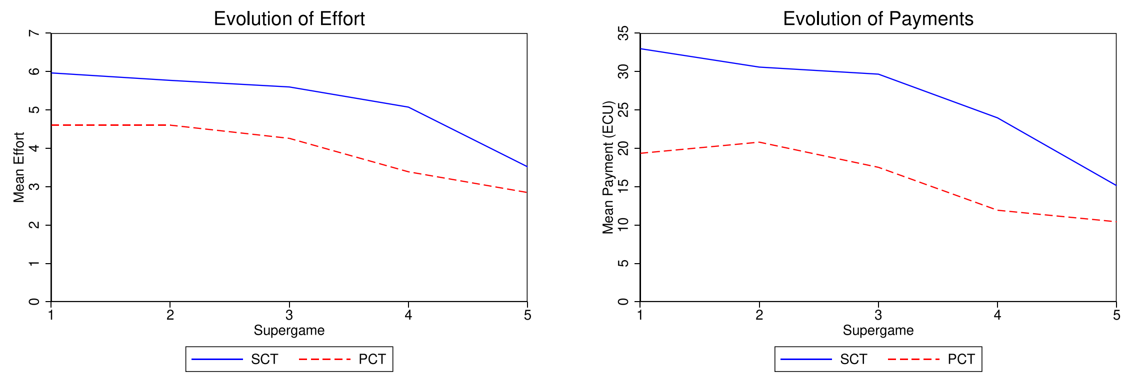
SCT and PCT do not differ in age (t-test: $p = 0.932$), gender (chi-square test: $p = 0.586$), or the number of boxes collected in the Bomb Risk task (t-test: $p = 0.888$). Treatment effects can therefore be attributed to the manipulation of the visibility of review thresholds.

Result 1: The Decline of Cooperation:

Public review commitment leads to a significant decline in cooperation, as predicted by the theoretical model. Fig. 1 compares the empirical cumulative distribution functions (ECDFs) of Agent's effort (left panel) and Principal's payment (right panel) across treatments. The PCT ECDFs lie above the SCT ECDFs for both effort and payments. This pattern indicates that low effort and low payments are more common under public commitment.

[2] The translated PCT instructions, with the corresponding SCT difference noted, are reproduced in Online Appendix B. The comprehension quiz is reproduced in Online Appendix C.

[3] This is approximately US$11.50 and corresponds to an hourly earnings rate about three times the local hourly minimum wage.



(a) Effort Evolution

(b) Payment Evolution

Fig. 2. Mean effort and payments by supergame and treatment.

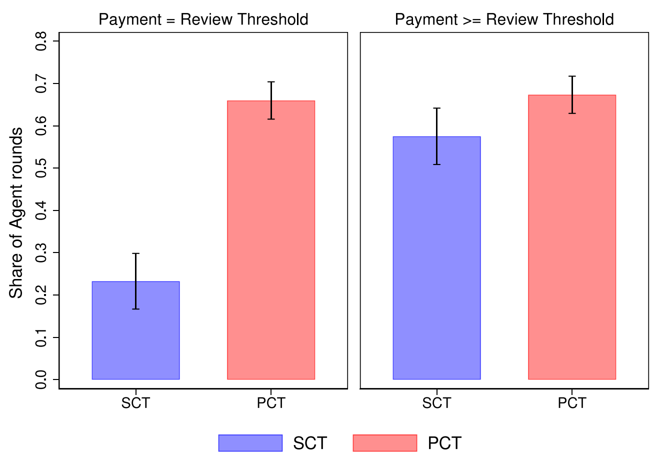


Fig. 3. Review Threshold Enforcement by Treatment. Capped lines show 95% confidence intervals.

Average Agent-level effort drops by 24%, from 4.91 units in the SCT to 3.74 units in the PCT (Mann–Whitney $z = 3.289, p = 0.001$); average Principal-level payments fall by 40%, from 24.57 to 14.85 ECU (Mann–Whitney $z = 4.558, p < 0.001$).[4] When review strategies are weaponized as public action-contingent threats, coordinated punishment sustains cooperation less effectively: trust and trade decline significantly.

Fig. 2 shows the evolution of average effort and payments across the five supergames. In the first two supergames, average Agent-level effort is significantly lower in PCT than SCT (4.60 versus 5.90; Mann–Whitney $z = 3.468, p < 0.001$). The effort gap remains significant in the final two supergames (3.05 versus 4.12; $z = 2.830, p = 0.005$). Payments are significantly lower in PCT in both the early and late periods.[5]

Agent effort also declines within both treatments (Wilcoxon signed-rank, $z = 5.216, p < 0.001$ for SCT; $z = 4.390, p < 0.001$ for PCT; Appendix Table A.3). In the final supergame, PCT effort reaches a low of 2.85 units versus 3.52 in SCT. Erosion characterizes both treatments, but public commitment yields a lower floor for productive effort, consistent with the predicted collapse of relational contracts under coordinated extortion.

[4] Detailed statistics are reported in Appendix Table A.2.

[5] See Appendix Table A.3 for details.

Result 2: Threshold Enforcement and Zero-Effort Extraction.

Public commitment makes the Agent's review threshold observable. We expected two natural consequences from this observability: Principals are more likely to meet visible thresholds, and some Agents use this visibility to obtain payments after providing no effort.

Fig. 3 reports two unconditional subject-level enforcement measures, each computed as a share of all Agent rounds. The first is the joint frequency with which a positive review threshold is posted and the Principal's payment exactly equals it (share of exact threshold payments); the second is the joint frequency with which a positive review threshold is posted and the payment weakly meets it (share of weak compliance). Both rise sharply under public commitment.

The joint share of exact threshold payments rises from 0.233 in SCT to 0.660 in PCT (Mann–Whitney $z = -6.295, p < 0.001$); the joint share of weak compliance rises from 0.575 to 0.673 ($z = -2.246, p = 0.025$). Public observability makes posted review thresholds salient and increases the likelihood that Principals meet them.[6]

Fig. 4 reports two unconditional subject-level shares, each computed out of all Agent rounds: the joint frequency of an Agent choosing zero effort and posting a positive review threshold, and the joint frequency of zero effort with a positive review threshold that the Principal's payment weakly meets. Both point estimates are higher in PCT: the zero-effort positive-threshold share is 0.113 in PCT versus 0.066 in SCT (Mann–Whitney $z = -1.239, p = 0.215$), while successful zero-effort extraction is 0.038 versus 0.007 ($z = -2.175, p = 0.030$). The positive-threshold difference is not statistically significant, but successful extraction is significantly more frequent under public commitment at the 5% level.

The extortion pattern is concentrated rather than universal. In SCT, 0 of 36 Agents extract successfully at zero effort in at least 10% of rounds; in PCT, 4 of 31 Agents do. At the stricter 25% cutoff, the corresponding counts are 0 of 36 in SCT and 1 of 31 in PCT (Appendix Table A.5). Public commitment therefore does not make every Agent an extortionist. It creates a strategic opportunity used by a meaningful minority, which is enough to distort the link between effort and payment.[7]

[6] The statistical tests are reported in Appendix Table A.4.

[7] Barron and Guo (2021, Section VI) develop an extension in which Agents face a private cost of committing to an extortionary threat, allowing cooperation and extortion to coexist in equilibrium. In our setting, this cost may be interpreted as a psychological cost of engaging in extortion. We attempted a structural estimation of this extension, but our data do not identify the cost distribution reliably enough to support parameter estimates or counterfactuals. We therefore do not report them.

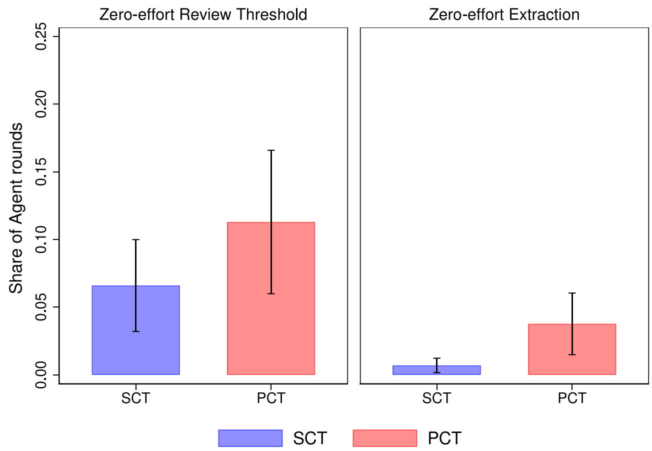


Fig. 4. Zero-Effort Positive Review Thresholds and Extraction by Treatment. Capped lines show 95% confidence intervals.

Table A.1
Summary statistics and treatment balance.

Variable	SCT (N = 72)	PCT (N = 62)	Overall (N = 134)	*p*-value
Age	21.94	21.92	21.93	0.932
	(1.54)	(1.87)	(1.69)	
Female	0.736	0.694	0.716	0.586
	(0.444)	(0.465)	(0.452)	
Risk Preference	49.03	48.73	48.89	0.888
(Boxes Collected)	(12.17)	(12.46)	(12.26)	

Notes: Standard deviations are in parentheses. The p-values for Age and Risk Preference are derived from two-sample t-tests with equal variances. The p-value for Female is derived from a Pearson χ^2 test. A Wilcoxon rank-sum test for Risk Preference yields consistent results ($p = 0.906$).

Table A.2
Treatment comparison: Effort and payments.

Variable	SCT (N = 36)	PCT (N = 31)	MW Test (z-stat)
Average Effort	4.91	3.74	3.289***
	(1.41)	(1.13)	
Average Payment	24.57	14.85	4.558***
	(9.20)	(5.19)	

Notes: The table reports role-restricted subject-level means with standard deviations in parentheses. Effort is calculated for Agents only; payment is calculated for Principals only. MW Test refers to the Wilcoxon rank-sum (Mann–Whitney) test comparing the two treatments. ** denotes $p < 0.05$; *** denotes $p < 0.01$.

5. Conclusion

Reputation and punishment sustain cooperation in decentralized exchange: information about past behavior disciplines opportunism (Kandori, 1992), and sanctions support self-governance (Ostrom et al., 1992; Fehr and Gächter, 2000). This paper asks when the same machinery can be turned against cooperation.

We provide an experimental test. The design builds on Barron and Guo (2021)'s insight that collective punishment becomes harmful when agents commit to observable threats, and on evidence that punishment can be misdirected or antisocial (Herrmann et al., 2008; Nikiforakis, 2008). Our contribution isolates one channel: public visibility turns a binding but hidden review threshold into a visible, more enforceable threat.

The results support this mechanism. Public commitment sharply reduces effort and payments while making the Agent's review threshold visible and more enforceable; consequently, Principals are more likely to meet the review threshold. Positive review thresholds after zero effort are higher descriptively under public commitment, and successful extraction is significantly more frequent. The evidence therefore supports both stronger threshold enforcement and a concentrated increase in realized zero-effort extraction.

These results complement Wilson and Vespa (2020), who study a repeated sender–receiver game in which efficient information transmission is supportable but often fails because players do not coordinate on how to divide the gains. When participants coordinate on compensating the sender with an information rent, honest revelation emerges. In our setting, by contrast, making the Agent's payment condition observable strengthens the Agent's distributional leverage and reduces productive cooperation. Taken together, the two studies suggest that outcomes in asymmetric repeated games depend not only on whether participants coordinate, but also on how the gains from cooperation are divided.

The broader implication is that transparency in reputation systems is not unambiguously beneficial. Observable threats can improve discipline, but can also make review thresholds credible and enforceable, especially when reports are hard to verify or can intimidate partners (Chassang and Padró i Miquel, 2019). Reputation systems thus require safeguards not only against moral hazard, but also against the strategic use of public feedback rules as leverage (Bolton et al., 2013). Coordinated punishment sustains cooperation only when the institutions organizing it also limit its misuse.

Appendix A. Statistical tables

See Tables A.1–A.5.

Appendix B. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.econlet.2026.113227.

Data availability

Data will be made available on request.

Table A.3
Dynamic erosion of cooperation: Early vs. late supergames.

	Early (SG 1–2)		Late (SG 4–5)		Within-Treatment Decline	
Variable	SCT	PCT	SCT	PCT	SCT Sign-Rank (z)	PCT Sign-Rank (z)
Average Effort	5.90	4.60	4.12	3.05	5.216***	4.390***
Average Payment	32.17	19.83	18.53	11.00	5.090***	4.742***
Between-Treatment Effect						
MW Test (z-stat)	Effort: 3.468***		Effort: 2.830***			
	Payment: 5.175***		Payment: 3.081***			

Notes: Role-restricted subject-level means are reported for the early (Supergames 1-2) and late (Supergames 4-5) periods. Effort tests use Agent subject-level means; payment tests use Principal subject-level means. The "Within-Treatment Decline" panel reports the Wilcoxon signed-rank z-statistics testing the drop in effort and payment from the early to late period within each treatment. The "Between-Treatment Effect" panel reports the Wilcoxon rank-sum (Mann–Whitney) z-statistics testing the difference between SCT and PCT in each respective period. ** denotes $p < 0.05$; *** denotes $p < 0.01$.

Table A.4
Mechanism tests: Review threshold enforcement and zero-effort extraction.

Variable	SCT (N = 36)	PCT (N = 31)	MW Test (z-stat)
Payment = review threshold	0.233	0.660	−6.295***
	(0.195)	(0.120)	
Payment ≥ review threshold	0.575	0.673	−2.246**
	(0.197)	(0.120)	
Zero-effort positive review threshold	0.066	0.113	−1.239
	(0.100)	(0.144)	
Zero-effort extraction	0.007	0.038	−2.175**
	(0.016)	(0.062)	

Notes: Subject-level means are reported for the Agent role only and are unconditional shares out of all Agent rounds. Standard deviations are in parentheses. A review threshold means $\hat{s} > 0$. For the payment rows, the event also requires that a positive review threshold is posted; for zero-effort extraction, the event requires $e = 0$, $\hat{s} > 0$, and $s \geq \hat{s}$. MW Test refers to the Wilcoxon rank-sum (Mann–Whitney) test comparing the two treatments. ** denotes $p < 0.05$; *** denotes $p < 0.01$.

Table A.5
Subject-level heterogeneity in extraction.

Type cutoff	SCT (N = 36)	PCT (N = 31)
Zero-effort extraction in at least 10% of rounds	0	4
Zero-effort extraction in at least 25% of rounds	0	1

Notes: The table reports counts of Agents whose unconditional subject-level extraction rate, computed out of all Agent rounds, exceeds the listed cutoff. Zero-effort extraction requires $e = 0$, $\hat{s} > 0$, and $s \geq \hat{s}$.

References

Barron, Daniel, Guo, Yingni, 2021. The use and misuse of coordinated punishments. Q. J. Econ. 136 (1), 471–504.

Bolton, Gary, Greiner, Ben, Ockenfels, Axel, 2013. Engineering trust: Reciprocity in the production of reputation information. Manag. Sci. 59 (2), 265–285. http://dx.doi.org/10.1287/mnsc.1120.1609.

Brown, Martin, Falk, Armin, Fehr, Ernst, 2004. Relational contracts and the nature of market interactions. Econometrica 72 (3), 747–780.

Chassang, Sylvain, Padró i Miquel, Gerard, 2019. Crime, intimidation, and whistle-blowing: A theory of inference from unverifiable reports. Rev. Econ. Stud. 86 (6), 2530–2553. http://dx.doi.org/10.1093/restud/rdy075.

Chen, Daniel L., Schonger, Martin, Wickens, Chris, 2016. oTree—An open-source platform for laboratory, online, and field experiments. J. Behav. Exp. Financ. 9, 88–97. http://dx.doi.org/10.1016/j.jbef.2015.12.001.

Conti, Allie, 2019. I accidentally uncovered a nationwide scam on Airbnb. VICE, Available at vice.com. (Accessed 1 September 2023).

Crosetto, Paolo, Filippin, Antonio, 2013. The "bomb" risk elicitation task. J. Risk Uncertain. 47 (1), 31–65. http://dx.doi.org/10.1007/s11166-013-9170-z.

Fehr, Ernst, Gächter, Simon, 2000. Cooperation and punishment in public goods experiments. Am. Econ. Rev. 90 (4), 980–994.

Fréchette, Guillaume R., Vespa, Emanuel, Yuksel, Sevgi, 2025. Repeated games. In: Snowberg, Erik, Yariv, Leeat (Eds.), In: Handbook of Experimental Methodology, vol. 1, Elsevier, pp. 197–234. http://dx.doi.org/10.1016/bs.hbem.2025.09.004.

Greif, Avner, Milgrom, Paul, Weingast, Barry R., 1994. Coordination, commitment, and enforcement: The case of the merchant guild. J. Political Econ. 102 (4), 745–776.

Herrmann, Benedikt, Thöni, Christian, Gächter, Simon, 2008. Antisocial punishment across societies. Science 319 (5868), 1362–1367.

Kandori, Michihiro, 1992. Social norms and community enforcement. Rev. Econ. Stud. 59 (1), 63–80. http://dx.doi.org/10.2307/2297925.

Klein, Tobias J., Lambertz, Christian, Stahl, Konrad O., 2016. Market transparency, adverse selection, and moral hazard. J. Political Econ. 124 (6), 1677–1713.

Krautmann, Anthony C., Sander, William, 1999. Grades and student evaluations of teachers. Econ. Educ. Rev. 18 (1), 59–63. http://dx.doi.org/10.1016/S0272-7757(98)00004-1.

Langbein, Laura, 2008. Management by results: Student evaluation of faculty teaching and the mis-measurement of performance. Econ. Educ. Rev. 27 (4), 417–428. http://dx.doi.org/10.1016/j.econedurev.2006.12.003.

Nikiforakis, Nikos, 2008. Punishment and counter-punishment in public good games: Can we really govern ourselves? J. Public Econ. 92 (1–2), 91–112.

Ostrom, Elinor, Walker, James, Gardner, Roy, 1992. Covenants with and without a sword: Self-governance is possible. Am. Political Sci. Rev. 86 (2), 404–417. http://dx.doi.org/10.2307/1964229.

Wilson, Alistair J., Vespa, Emanuel, 2020. Information transmission under the shadow of the future: An experiment. Am. Econ. J.: Microeconomics 12 (4), 75–98. http://dx.doi.org/10.1257/mic.20170403.