Journal of the Economic Science Association (2024) 10:113–135
https://doi.org/10.1007/s40881-023-00153-3

ORIGINAL PAPER

The seeds of success: the pivotal role of first round cooperation in public goods games

Mehmet Y. Gürdal[1] · Orhan Torul[1] · Mustafa Yahşi[2]

Received: 20 August 2022 / Revised: 21 October 2023 / Accepted: 23 October 2023 /
Published online: 7 January 2024

Abstract

This paper examines cooperation and punishment in a public goods game in Istanbul. Unlike prior within-subject designs, we use a between-subject design with separate no-punishment and punishment conditions. This approach reveals that punishment significantly increases contributions, demonstrating the detrimental effect of having prior experience without sanctions. We highlight two critical factors—heterogeneous initial contributions across groups and how subjects update their contributions based on prior contributions and received punishment. An agent-based model verifies that the interaction between these two factors leads to a strong persistence of contributions over time. Analysis of related data from comparable cities shows similar patterns, suggesting our findings likely generalize if using a between-subject design. We conclude that overlooking within-group heterogeneity biases cross-society comparisons and subsequent policy implications.

Keywords Public goods experiment · Punishment · Cooperation · Culture

JEL Classification C91 · C92 · D81 · J16

✉ Orhan Torul
orhan.torul@boun.edu.tr

Mehmet Y. Gürdal
mehmet.gurdal@boun.edu.tr

Mustafa Yahşi
myahsi15@ku.edu.tr

[1] Department of Economics, Boğaziçi University, Bebek, 34342 Istanbul, Turkey

[2] Department of Economics, Koç University, Rumelifeneri Yolu, Sarıyer, 34450 Istanbul, Turkey

1 Introduction

Achieving cooperation in group settings where individual and collective interests conflict poses a longstanding challenge. Individuals are incentivized to act in their own self-interest rather than cooperating for the collective good, which can lead to detrimental outcomes like the tragedy of the commons. Effective cooperation requires aligning individual and group interests through mechanisms like incentives, penalties for defection, communication, trust-building, and reputation systems. However, this remains difficult, as individuals often find ways to free-ride if acting selfishly benefits them more than cooperating. Examples include overfishing, public goods provision, and reducing carbon emissions.

Public goods experiments using the voluntary contribution mechanism (*VCM*) are a classic way to study cooperation in these social dilemma situations. In a typical *VCM* experiment, subjects are matched in small groups and contribute towards a non-excludable public good over multiple periods. When the group size is n, and one unit contribution to the group account increases each member's earnings by p, with $p < 1$, and $np > 1$, the Nash equilibrium is zero contribution, whereas the efficient outcome is full contribution. Results from *VCM* experiments show that individuals generally start with some positive contribution, which decays over time without incentives (Isaac & Walker, 1988).

In their seminal studies, Fehr and Gächter (2000, 2002) implement a design where subjects play a game of known length, observe each others' contributions each period, and decide how much to deduct from the earnings of other group members, i.e., *sanction* each other.[1] This deduction is costly for the subject; therefore, it is classified as a form of strong reciprocity (*costly altruism*). When subjects are allowed to assign punishment points to their group members, this can facilitate cooperation and keep the contribution levels high in the experiment. Herrmann et al. (2008) extended this approach by conducting a study comparing the effectiveness of punishment in sustaining cooperation across 16 cities worldwide. They found that punishment successfully increased cooperation in Western cities like Boston, Copenhagen, St. Gallen, Zurich, and Nottingham.[2] However, it failed to do so in Istanbul and other similar cities. Unlike Western subjects,

[1] Earlier examples of similar games with sanctioning mechanisms include Yamagishi (1986), in which subjects could punish the least cooperative group member, and Ostrom et al. (1992), in which there is a fixed cost of sanctioning and being sanctioned. Because the game length is unknown to the subjects in the latter study, it is not possible to rule out reputational incentives behind contribution and sanctioning decisions.

[2] The average contribution rate was at least 75% in all of these cities when punishment was available.

Istanbul participants exhibited more antisocial punishment, where high contributors were targeted. Overall, punishment had low effectiveness in these cities.[3, 4]

Our study revisits cooperation and punishment in Istanbul using the same subject pool but a between-subject design, in contrast to the within-subject design used by Herrmann et al. (2008). In their study, all subjects first played the *VCM* experiment without punishment and then with punishment. In our study, subjects participated in either the game without punishment (N-experiment) or with punishment (P-Experiment), but not both. Under this between-subject approach, we find substantially higher contributions in Istanbul when punishment is introduced compared to no punishment. However, average contributions still remain below levels observed in Western cities.

We identify two critical factors behind cooperation outcomes in the P-Experiment:

(1) The distribution of first-period contributions exhibits significant heterogeneity across groups. Groups starting at high initial cooperation sustain those levels, while groups starting at low cooperation remain low throughout.
(2) Subjects follow simple contribution rules based on prior contributions and received punishment. We estimate these linear rules from the data.

The interaction of these two factors generates strong persistence in contributions over time. To demonstrate this, we conduct a counterfactual experiment using an agent-based model. Feeding the estimated decision rules and actual first-period data accurately reproduces the evolution of contributions over the course of the game.

Analyzing data from Herrmann et al. (2008), we find similar first-period heterogeneity and contribution persistence over time in cities resembling Istanbul. In particular, these stylized patterns are common across cities that share socio-economic similarities with Istanbul, particularly in Athens (Greece), Dnipropetrovsk (Ukraine), Samara (Russia), Minsk (Belarus), Riyadh (Saudi Arabia), and Muscat (Oman). Our Istanbul results likely extend to these cities under a between-subject design.

Our findings suggest that accounting for heterogeneous initial cooperation, rather than just city-level averages, is critical for valid cross-society comparisons. The decay induced by a prior N-experiment in a within-subject design masks actual cooperation capacity. More broadly, populations exhibiting diversity in initial cooperation may see larger gains from institutions designed to enhance cooperation. This study isolates the effect of experimental design and highlights the pivotal role

[3] Antisocial punishment occurs when a subject punishes a group member who contributed more than them to the group account. Herrmann et al. (2008) argue that the extent of antisocial punishment is negatively correlated with contribution levels, and subject pools where antisocial punishment is rare are more likely to achieve high contribution rates.

[4] Other studies also find location-specific differences in cooperation and punishment. For example, Gächter and Herrmann (2009) show higher cooperation and prosocial punishment in Switzerland than antisocial punishment in Russia. Another example is Buchan et al. (2011), who find that cooperation with a global group is positively correlated with a country's globalization level. For further work on location-specific differences, see Cárdenas et al. (2012) and Lamba and Mace (2011).

of initial conditions for cooperation outcomes. It provides guidance for designing future cross-society cooperation experiments to produce desirable results.

The rest of the paper is organized as follows: in Sect. 2, we describe the experiment. In Sect. 3, we report our empirical findings and discuss salient patterns that guide contribution and sanction decision rules. In Sect. 4, we propose and report the results of our agent-based model motivated by the estimated decision rules based on prior contributions and sanctions. In Sect. 5, we compare our findings in detail with those by Herrmann et al. (2008), and Sect. 6 concludes.

2 Experimental design and procedures

The experiment is based on the design by Fehr and Gächter (2000) and involves two treatments. In the *N-experiment*, subjects are randomly matched into groups of 4 and interact within the same group for 10 periods. Each period, subjects receive an endowment of 20 tokens, from which they can contribute to a "group project". For every token invested in the group project, each group member earns 0.4 tokens. Subjects' period earnings are calculated as the sum of earnings from the group project and the part of the endowment not invested in the group project. The *P-experiment* builds on the N-experiment but adds a punishment stage. After observing group members' contributions (but not their identities), subjects can assign costly punishment points. Each point assigned costs the subject 1 token and reduces the target's earnings by 3 tokens. The total reduction in a subject's earnings is limited to the earnings from the contribution stage. We discuss further details of the two stages in the Appendix.

3 Results

We provide summary statistics for the observed contribution levels in Table 1. The initial contributions in our study are similar in both treatments, starting around 9 tokens. However, by the 10th period, average contributions decline to 2.85 tokens in the N-experiment, whereas they rise moderately to about 12 tokens in the P-experiment. Overall, contributions in the P-experiment are significantly higher than in the N-experiment ($p = 0.019$), even though contributions in the first period do not differ significantly between the two experiments ($p = 0.950$). Non-parametric testing reveals sufficient statistical power ($pw = 0.78$) to detect significant differences in average contributions between the N and P-experiments across all periods.[5]

Herrmann et al. (2008) employ the design by Fehr and Gächter (2000) to measure the performance of the punishment mechanism among 1120 subjects in 16 different

[5] Since the power value of the test is reasonably high at 0.78, it is safe to assume that there is a significant difference in the average contributions between the P-experiment and the N-experiment with our sample sizes. For a detailed discussion on statistical power analysis, see Cohen (1988).

Table 1 Mean contributions and earnings

	Contribution in period 1			Contribution in all periods			Earnings	
	N	P	*p* value	N	P	*p* value	N	P
Istanbul—this study	8.9	9.1	0.950	6.1	10.9	0.019	23.6	19.4
Istanbul—Herrmann et al. (2008)	8.9	6.5	0.034	5.4	7.1	0.326	23.3	17.0
Boston—Herrmann et al. (2008)	13.0	16.0	0.012	9.3	18.0	0.002	25.6	27.9
CopenhagenHerrmann et al. (2008)	14.1	15.4	0.088	11.5	17.7	0.001	26.9	27.7

This table presents the average contributions (out of a total of 20 tokens) and the corresponding *p* values from Mann–Whitney tests. These tests consider the group average contributions as independent observations and examine whether there is a significant difference in contributions between the N-experiment and the P-experiment

cities worldwide, including Istanbul. We report the contributions of subjects in Istanbul, Boston, and Copenhagen from that study in Table 1.[6]

Using data from Istanbul, Herrmann et al. (2008) find a difference between the two treatments in Period 1, with lower contributions in the P-experiment than the N-experiment. However, they find no significant difference when considering all periods. Our summary statistics and non-parametric test results suggest that a between-subject design increases the effectiveness of punishment in sustaining cooperation in Istanbul. However, compared to results from Boston and Copenhagen, subjects in Istanbul still contribute at lower rates on average, even with punishment. Additionally, compared to those cities, the availability of punishment leads to lower average earnings in Istanbul, both in the current study and Herrmann et al. (2008).

We next report and discuss the results from our two treatments.

3.1 N-experiment

Contributions in the N-experiment exhibit the typical decay pattern seen in public good games. We present the evolution of contributions averaged over groups for the N and P-experiments in Fig. 1. The frequency of zero contributions in the N-experiment is 27% in the first half but increases to 46% in the second half. Of the 15 groups, 11 end up at very low contribution levels (3 tokens or less), 3 at moderate levels (6.75–8.5 tokens), and only 1 reaches 12 tokens by period 10. While these values point out some degree of heterogeneity, the dominant trend in the N-experiment

[6] The latter two cities are where the highest average contributions were reported in the P-experiment and the N-experiment, respectively. We emphasize that we use a between-subject design, whereas Herrmann et al. (2008) use a within-subject design, in which the P-experiment follows the N-experiment.

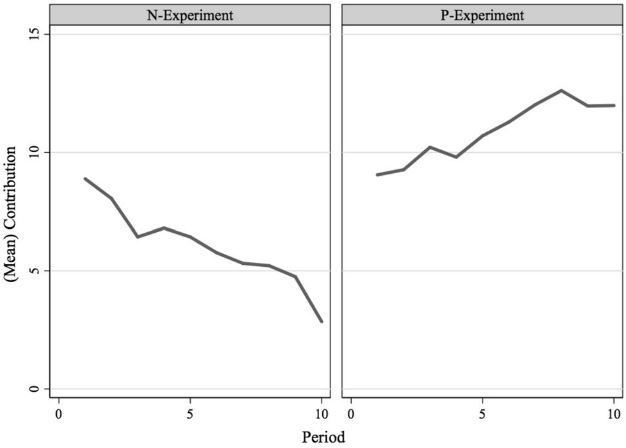

Fig. 1 Timeline of average contribution by treatment. *Notes:* The left and right panel displays the average contribution in the N-experiment and P-experiment, respectively

is a clear decline in contributions over time. This contrasts the evolution of contributions in the P-Experiment, as discussed in detail in the next Sect. 3.2.[7]

3.2 P-experiment

The P-experiment exhibits two salient patterns based on initial average contributions within each group, as illustrated in Fig. 2. First, we observe that group member update their contributions towards the group average. When a group member observes that her contribution falls short of the group average, she seldom decreases her contribution in the next period. Similarly, when her contribution exceeds the group average, next-period contribution is frequently either lower or the same. Second, we observe that punishment decisions are often social. Subjects are more likely to punish those who contribute less than them, with the likelihood and amount of punishment increasing as the difference in contributions grows. Further, the amount of punishment is negatively related to the contribution of the target subject and positively related to the average contribution of the remaining group members. We report the quantitative details of these empirical patterns in the Appendix.

We next propose an agent-based model that combines the salient empirical decision patterns of our experiment with first-period contributions.

[7] In the first period, the mean (and standard deviation) of average contributions in the N-Experiment was 8.88 (6.66), and in the P-Experiment, it was 9.05 (6.49). By the tenth period, these statistics had changed significantly, with the mean contribution in the N-experiment decreasing to 2.85 (5.42) and the mean contribution in the P-experiment increasing to 11.98 (6.77).

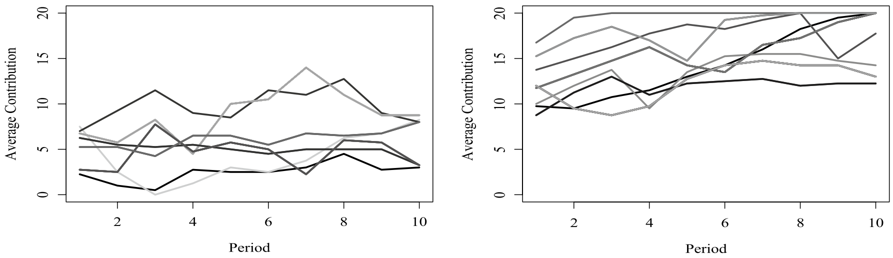

Fig. 2 Evolution of average contributions in the P-experiment. *Notes:* The left panel shows the average contribution over 10 periods for the 7 groups with the lowest first-period average contributions, while the right panel shows the same statistic for the remaining 8 groups

4 Agent-based modeling

In addition to the between-subject design, initial heterogeneity across groups emerges as a decisive factor affecting cooperation in the P-experiment.[8] Our findings show that if subjects in Istanbul start off collaborating, they sustain high contribution levels comparable to Boston and Copenhagen. Figure 2, which clusters groups by their average contribution in the initial round, highlights the importance of the first-period contributions. The seven groups with low starting contributions are on the left, while the eight groups with high starting contributions are on the right. Figure 2 reveals that groups that contribute above 43.75% on average in the first round maintain rates of at least 61.25% by the final period.

To demonstrate the decisive role of initial contributions, we use an agent-based modeling approach. We first take the first-period contributions data from the P-experiment. Then we impose simple state-dependent linear decision rules based on patterns observed in the data. These rules determine how agents update contributions over time. We then run Monte Carlo simulations for each of the 15 groups and examine the evolution of contribution over periods. The details of our procedure are discussed in the Appendix and the timeline of the agent-based modeling algorithm is summarized in Fig. 3.

Figure 4 displays the results of the simulations. For each group, it shows the average contributions from the simulations (dotted lines) with 2-standard deviation confidence intervals (shaded areas). These are plotted alongside the actual average contributions from the experiment (solid lines). Despite its simplicity, the imposed decision rule mimics not only the actual final contributions but also the contribution dynamics for most groups.

[8] Specifically, we argue that drawing conclusions about the success rates of cities by focusing only on group averages, as in Herrmann et al. (2008), can be misleading, as this approach obscures the decisive role of first-round results through aggregation.

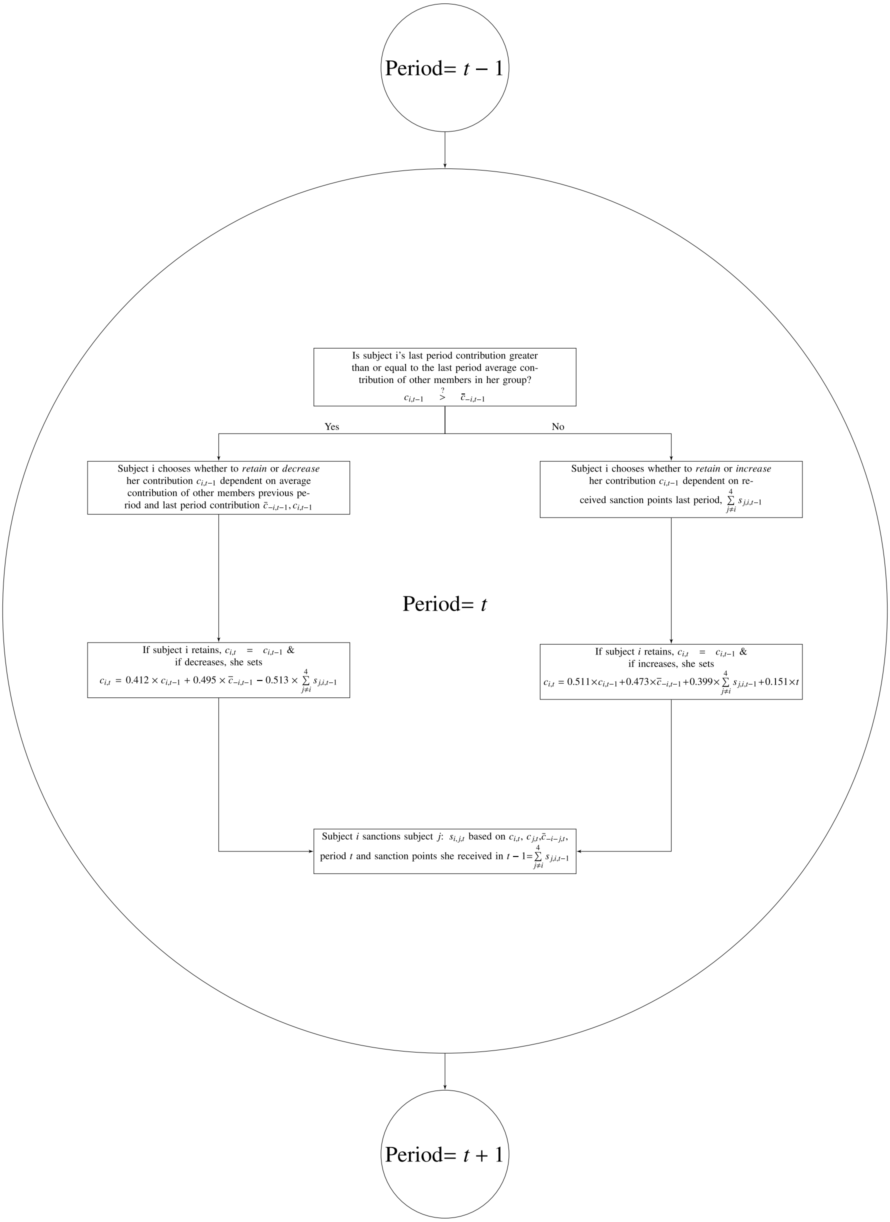
Period= $t-1$
Is subject i's last period contribution greater than or equal to the last period average contribution of other members in her group? $c_{i,t-1} \overset{?}{>} \bar{c}_{-i,t-1}$
Yes
No
Subject i chooses whether to *retain* or *decrease* her contribution $c_{i,t-1}$ dependent on average contribution of other members previous period and last period contribution $\bar{c}_{-i,t-1}, c_{i,t-1}$
Subject i chooses whether to *retain* or *increase* her contribution $c_{i,t-1}$ dependent on received sanction points last period, $\sum_{j\neq i}^{4} s_{j,i,t-1}$
Period= t
If subject i retains, $c_{i,t} = c_{i,t-1}$ & if decreases, she sets $c_{i,t} = 0.412 \times c_{i,t-1} + 0.495 \times \bar{c}_{-i,t-1} - 0.513 \times \sum_{j\neq i}^{4} s_{j,i,t-1}$
If subject i retains, $c_{i,t} = c_{i,t-1}$ & if increases, she sets $c_{i,t} = 0.511 \times c_{i,t-1} + 0.473 \times \bar{c}_{-i,t-1} + 0.399 \times \sum_{j\neq i}^{4} s_{j,i,t-1} + 0.151 \times t$
Subject i sanctions subject j: $s_{i,j,t}$ based on $c_{i,t}$, $c_{j,t}$,$\bar{c}_{-i-j,t}$, period t and sanction points she received in $t-1$=$\sum_{j\neq i}^{4} s_{j,i,t-1}$
Period= $t+1$

Fig. 3 Timeline of the model. *Notes*: Fig. 3 displays the timeline of agent-based modeling for the P-experiment. Only first-period contributions are taken from the experiment, while the remaining data are generated through agent-based modeling. Subjects' decisions to keep or change their contribution level in subsequent periods are determined by Table 3. If a subject changes their contribution level, their contribution in the next period is generated using Table 4. Sanctioning decisions are made using Table 6

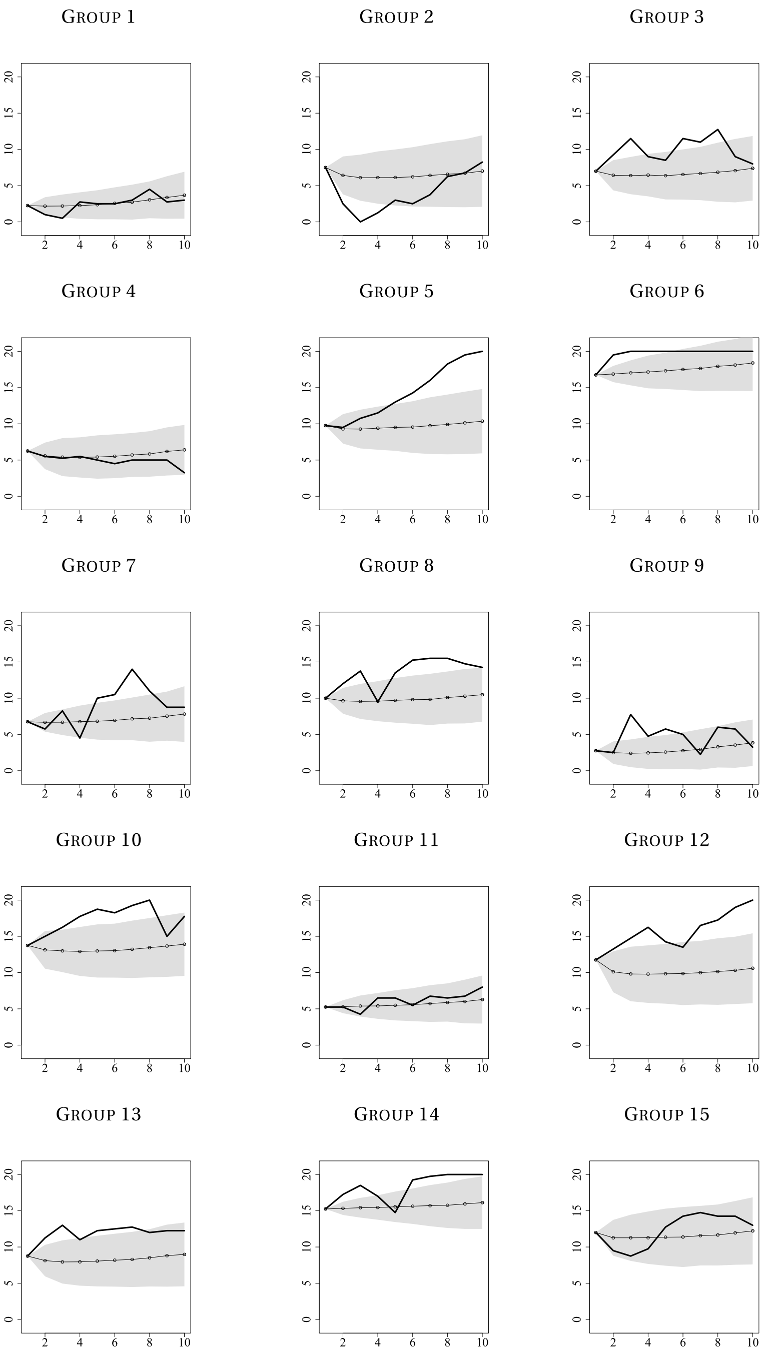

Fig. 4 Agent-based model simulation results. *Notes:* The horizontal axis shows the periods, and the vertical axis shows the average group contribution. The solid line represents the group averages from the P-experiment data. The dotted line represents the group-specific average agent-based model results from the Monte Carlo simulations, with the shaded gray areas representing the resulting 2-standard deviation confidence intervals

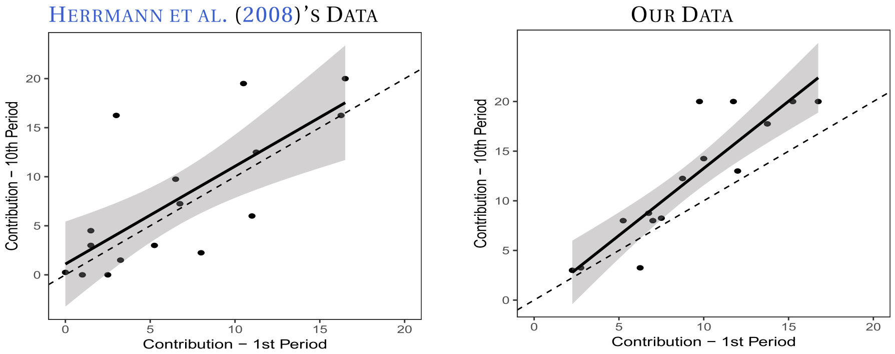

Fig. 5 Comparison of average contributions in the P-experiment. *Notes:* Each point in the figure represents the average contribution of a group in the P-experiment. The horizontal axis shows the group's average contribution in the first period, and the vertical axis shows the group's average contribution in the tenth period. Groups below the dashed 45° line did not improve their average contribution from the first period, while groups above the line did improve their average contribution. The solid line represents the linear best-fit line, and the shaded gray areas show the 95% confidence intervals

5 Comparison with Herrmann et al. (2008)

The heterogeneity observed in the initial average contributions and its persistence throughout the game is not unique to our experiment. As such, the detrimental effects of a within-subject design may extend beyond Istanbul.

To compare our findings to those of Herrmann et al. (2008) for Istanbul, we plot first-period average group contributions along with their last-period counterparts in Fig. 5. This illustration shows that both datasets exhibit a high degree of variability in the first-period average group contributions.

However, there are two stark differences: First, only 2 out of 15 groups in our experiment contributed at a rate lower than 25% in the first period, compared to 7 out of 16 groups in Herrmann et al. (2008)'s experiment. Second, while the slope between the first and last period average group contributions in Herrmann et al. (2008)'s data are close to unity (0.997, with a standard error of 0.243), it is considerably steeper in our experiment (1.352, with a standard error of 0.186). This suggests that while average group contributions in Herrmann et al. (2008) stagnated on average over time, groups in our experiment managed on average to raise their average group contributions over periods.

Given that our experiment was conducted at the same university with the same subject pool, we argue that these differences are due to our between-subject design versus Herrmann et al. (2008)'s within-subject design, in which all average group contributions converged to zero by the last period of the no-punishment treatment.

Istanbul was not the only city that demonstrated significant heterogeneity in average first-period contributions that persisted throughout the P-experiment. The experiments by Herrmann et al. (2008) were conducted in various cities with marked socio-economic differences. To explore which of these cities resemble Istanbul the most, we conduct a principal component analysis (PCA) using three variables: (i) straight-line physical distance to Istanbul, (ii) GDP per capita (in 2017 current US dollars), and (iii) cultural and psychological distance to Istanbul (Turkey) via Muthukrishna (2018)'s WEIRD scale index. Our PCA reveals that a cluster of six cities from Herrmann et al. (2008) resembles Istanbul the most: Athens (Greece), Dnipropetrovsk (Ukraine), Samara (Russia), Minsk (Belarus), Riyadh (Saudi Arabia), and Muscat (Oman).[9]

Figure 6 shows the first- and last-period average group contributions for the six cities from Herrmann et al. (2008) that most resemble Istanbul. All of these cities exhibit significant first- and last-period heterogeneity in average group contributions, with a strong positive correlation between the two that is close to unity. Additionally, a city-fixed-effect regression of last-period average group contributions on first-period average group contributions for these six cities yields a slope coefficient of 0.975 (with standard error 0.140), which is not statistically different from the slope of 0.997 (with standard error 0.243) that we estimate for Istanbul.[10] Therefore, we conclude that Istanbul does not single out with its idiosyncrasy vis-à-vis heterogeneity in first-period average contributions and its persistence throughout the game. Instead, these patterns are statistically common across cities that share socioeconomic similarities with Istanbul. As such, the increase in Istanbul's contribution rates due to the between-subject design could plausibly extend to other cities with similar characteristics.

6 Discussion and concluding remarks

In this paper, we show that contributions in a public goods game with punishment are significantly higher in Istanbul under a between-subject design than under a within-subject design in which the no-punishment condition precedes the punishment condition. This highlights the detrimental effect of prior experience without punishment on cooperation. However, Istanbul's average contribution remains below Western city levels, suggesting that limited cooperation persists even with sanctions.

We identify two key factors behind cooperation patterns: heterogeneous initial contributions, extending Burlando and Guala (2005), and simple contribution updating rules based on prior contributions and sanction points estimated from the data. An agent-based model verifies that the interaction of these factors generates strong persistence in contribution levels over time.[11]

[9] Four of the cities in Herrmann et al. (2008) do not have scores on Muthukrishna (2018)'s WEIRD scale index. We imputed values for these cities as follows: for Athens of Greece, we used Cyprus as a proxy; for Riyadh of Saudi Arabia and Muscat of Oman, we used neighboring Gulf countries as proxies; and for Copenhagen of Denmark, we used Sweden and Norway as proxies. The resulting Fig. 7 is in the Appendix. Details of our PCA exercise are available upon request.

[10] The city fixed-effect regression of the six cities yields a constant coefficient of −1.139 with a standard error of 1.467, while the constant coefficient for Istanbul is 1.100 with a standard error of 2.022. The corresponding Fig. 8 is in the Appendix.

[11] See also Andreoni and Miller (2002), who show that altruistic behavior is rationalizable, consistent, yet highly heterogeneous across individuals.

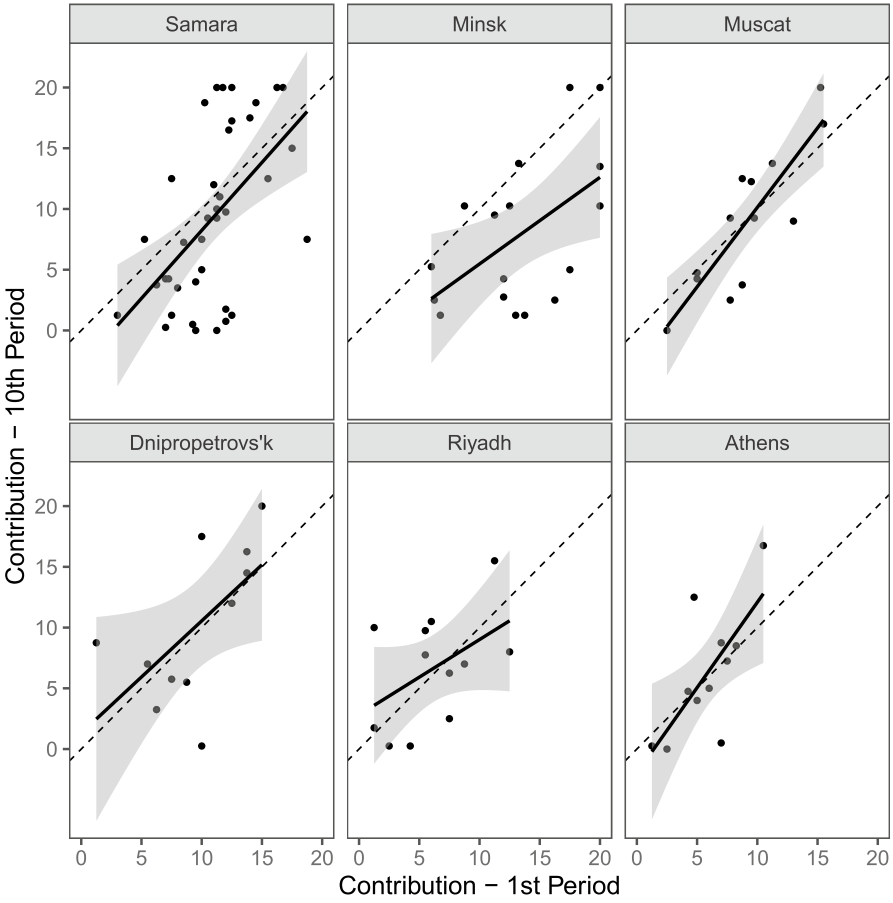

Fig. 6 First-period and last-period contributions by city (Herrmann et al., 2008). *Notes:* Each point in the figure represents the average contribution of a group in the P-experiment. The horizontal axis shows the group's average contribution in the first period, and the vertical axis shows the group's average contribution in the tenth period. Groups below the dashed 45-degree line did not improve their average contribution from the first period, while groups above the line did improve their average contribution. The solid line represents the linear best-fit line, and the shaded gray areas show the 95% confidence intervals

The data analysis from Herrmann et al. (2008) reveals similar heterogeneity and persistence in contribution levels over time in cities that resemble Istanbul. Therefore, our results are likely generalizable to these settings if a between-subject design is employed, which would eliminate the cooperation decay induced by a prior no-punishment condition.

Our results highlight the crucial role of initial contributions in shaping subsequent cooperation within a group, which is consistent with studies that have shown the effectiveness of grouping subjects based on contribution levels. For example, Gunnthorsdottir et al. (2007) find that contribution decay is lower when subjects are matched based on prior actions rather than randomly matched. Other relevant studies include Gächter and Thöni (2005), Ones and Putterman (2007), and Gunnthorsdottir et al. (2010). Brekke et al. (2011) also allow for endogenous group formation based on charity donations and find that cooperation is improved among donors.

Similar to Gächter et al. (2010), we posit that culture affects cooperation through beliefs and punishment responses. The variance we observe in initial contributions underscores the power of beliefs, which is consistent with the reduced contributions observed following a within-subject design as in Herrmann et al. (2008). Personal risk preferences also influence first-round contributions, which are then adjusted based on received sanctions. A group that starts with low contributions and antisocial punishment risks cooperation failure.

Our study makes several contributions. It demonstrates the critical impact of experimental design and initial conditions on cooperation outcomes. It provides guidance for robust cross-society experiment design by underscoring within-group heterogeneity. Our findings uniquely isolate the effect of first-round divergence, complementing research on culture and conditional cooperation in social dilemmas. The insights into contribution updating rules and belief formation advance theoretical understanding of how cooperation evolves.

Future work should further explore the sources of heterogeneous initial contributions and beliefs across individuals. Overall, highlighting within-society variation rather than just cross-society differences is critical for drawing valid inferences and crafting policies to encourage cooperation. This study demonstrates the inadequacy of only considering city-level aggregates when cooperation hinges fundamentally on initial beliefs within subgroups.

Appendix

Details of the experiments

In this appendix, we discuss the details of our experiments.

The contribution stage

In the contribution stage, subjects are allocated to groups of four and remain in the same group for 10 periods. In each period, subjects are given an initial endowment of 20 tokens and simultaneously choose how much to invest in the group project. For each group, the sum of all contributions is multiplied by 0.4 and then returned to each group member.

If g_i is the individual contribution made by player i and G is the total contribution made to the group project in player i's group, then the total payoff in token terms for player i in the N-experiment is given by:

$$\pi_i = 20 - g_i + 0.4G \tag{1}$$

After the contribution stage, subjects are informed of their individual contributions and the total payoff for all members of their group without disclosing the identity of the subjects. Note that the contribution stage is the only stage of the N-experiment and is repeated for 10 periods in both treatments.

The punishment stage

The punishment stage takes place only in the P-experiment. After the contribution stage, all subjects simultaneously decide whether to punish other members of their group. For each token player i assigns to player j, player i loses 1 token, and player j loses 3 tokens.

Let p_{ij} denote the punishment tokens assigned by player i to player j of the same group. If $p_i = \sum_{j \neq i} p_{ij}$ is the total punishment points assigned by player i to other group members and $p_{-i} = \sum_{j \neq i} p_{ji}$ is the total punishment points assigned to player i by other group members, then the total payoff in token terms for player i in the P-experiment is given by:

$$\pi_i = \max\left\{0, (20 - g_i + 0.4G - p_i - 3p_{-i})\right\}. \qquad (2)$$

After the punishment stage, subjects are informed of their individual contributions, the punishment tokens they sent and received, and the total payoffs of all members of their group without disclosing the identity of the subjects. The punishment stage is repeated for 10 periods after the contribution stage in the P-experiment.

Experimental procedures

The experiment was conducted with 120 subjects in 10 sessions at the Economics Laboratory of Boğaziçi University in Istanbul. Each session involved 12 subjects. We used a between-subject design, with 60 subjects participating in the N-experiment only and the other 60 subjects participating in the P-experiment only.[12] An email was sent to subjects who had previously expressed interest in participating in economics experiments. Subjects could register online for a date and time of their choosing. No subject participated more than once, and the sessions lasted an average of 45 min. Subjects were paid in cash at the end of the experiment at an exchange rate of 0.1 Turkish liras per token.[13]

[12] Note that our study differs from Herrmann et al. (2008), who used a within-subject design for the majority of their subject pools, including Istanbul. In their design, the P-experiment followed the N-experiment, while in ours, subjects participated in only one treatment. Other differences between our experiment and theirs include the average session size (21 in Herrmann et al. (2008) and 12 in ours) and the total number of subjects (64 for Istanbul in Herrmann et al. (2008) and 60 in ours).

[13] At the time of the experiments (November 9–11, 2015), 1 Turkish lira corresponded to approximately $0.35 United States dollars. The interface our subjects encountered was the Turkish translation of the one by Herrmann et al. (2008).

Empirical patterns in the P-experiment

In this appendix, we quantify the salient patterns in the P-experiment. One decisive pattern is that the way group members update their contributions in the P-experiment is closely related to the respective group average. Specifically, when a group member observes that their contribution falls short of the group average, they rarely decrease their contribution in the next period. Instead, they almost always either increase their contribution or keep it the same. Similarly, when their contribution exceeds the group average, their next period's contribution is often either lower or the same. We report the frequencies of the direction of change in contributions conditional on how one's last period contribution compares to the group average in Table 2.

We observe further salient patterns from the P-experiment. One decisive one is that the way group members update their contributions is closely related to the respective group average. Specifically, when a group member observes that their contribution falls short of the group average, they rarely decrease their contribution in the next period. Instead, they almost always either increase their contribution or keep it the same. Similarly, when their contribution exceeds the group average, their next period's contribution is often either lower or the same. We report the frequencies of the direction of change in contributions conditional on how one's last period contribution compares to the group average in Table 2.

Decision to change contributions

We run a series of logit and *OLS* regressions to unveil the driving forces behind the contribution decisions.

In a set of logit regressions reported in Table 3, we first examine the factors that influence decisions to revise (increase or decrease) contributions in the P-experiment. We also control for certain attitudinal and demographic information about subjects, as well as group-level fixed effects.[14]

In Models 1 and 2, the dependent variable, *change*, takes the value 1 if a subject's contribution in the current period differs from their contribution in the previous period and 0 otherwise. Our estimates reveal that the only variable significantly affecting the *change* variable is the number of punishment points received from other group members in the previous period (Model 1). This effect is robust to controlling for other factors (Model 2).

[14] The control variables used in the regressions are: age (in years), gender (dummy), number of older siblings, number of younger siblings, only child status (dummy), trust in people (binary, based on the question "Generally speaking, would you say that most people can be trusted or that you need to be very careful in dealing with people?"), risk aversion (Likert scale, based on the question "How willing are you to take risks in general?"), number of economics classes taken (censored at 4), number of friends in the session, membership in an organization (indicator variable), and reliability of responses in the experiment (Likert scale).

Table 2 Contribution change by group average

Group average	Next period contribution		
	Increases	Remains the same	Decreases
Lower	39 (16.5%)	81 (34.3%)	116 (49.2%)
Equal	9 (13.4%)	53 (79.1%)	5 (7.5%)
Higher	161 (67.9%)	45 (19.0%)	31 (13.1%)

This table shows the frequency of next-period contribution changes by subjects in the P-experiment based on the comparison of their current-period contribution to the group average. Percentages are calculated by row and represent the proportion of subjects in each category

In Models 3 and 4, the dependent variable, *increase*, takes the value 1 if a subject's contribution in the current period is greater than their contribution in the previous period and 0 otherwise. In these models, we restrict our working sample to observations featuring only subjects contributing less than the previous period's group average. Once again, the only variable with a significant effect is the number of punishment points received from other group members in the previous period (Model 3). This effect is also robust to controlling for other factors (Model 4).

In Models 5 and 6, the dependent variable, *decrease*, takes the value 1 if a subject's contribution in the current period is less than their contribution in the previous period and 0 otherwise. In these models, we again restrict our working sample to a subset of observations, this time featuring only subjects contributing more than or equal to the previous period's group average. In these estimates, the average contribution of other group members in the previous period and the subject's own contribution in the previous period significantly affect their contribution in the next period (Model 5). When controlling for other factors (Model 6), punishment points received from other group members in the previous period also have a significant effect.

After investigating the conditional frequencies of change decisions, we next examine the intensive margin, i.e., the magnitude of changes in contributions due to these factors.

Magnitude of change in contributions

We run a series of logistic and ordinary least squares (*OLS*) regressions to identify the factors that drive contribution decisions. Our findings are reported in Table 4. In this table, we use *contribution* as the dependent variable and restrict our working sample to observations featuring subjects who revise their contribution in the current period. Models 1 and 2 include all subjects who either increased or decreased their contributions (but not those who kept it the same). The number of punishment points received from other group members in the previous period, the average contribution of other group members in the previous period, and the previous period's contribution all have a significant effect on the next period's contributions (Model 1). These effects remain significant when controlling for other factors (Model 2).

Table 3 Determinants of contributions changes (LOGIT)

	(1)	(2)	(3)	(4)	(5)	(6)
Received points in $t-1$	0.314**	0.249**	0.175*	0.214**	0.111	0.190**
	(0.124)	(0.108)	(0.100)	(0.093)	(0.076)	(0.089)
Other's average contribution in $t-1$	−0.047	−0.047	0.043	0.057	−0.318***	−0.342***
	(0.044)	(0.038)	(0.064)	(0.047)	(0.066)	(0.059)
Contribution in $t-1$	0.017	−0.005	0.035	−0.010	0.216***	0.265***
	(0.029)	(0.037)	(0.077)	(0.060)	(0.040)	(0.046)
Period	0.084	−0.020	−0.068	−0.006	−0.018	−0.039
	(0.072)	(0.055)	(0.054)	(0.075)	(0.067)	(0.078)
Final period	−0.395	−0.067	−0.034	−0.227	0.376	0.392
	(0.288)	(0.242)	(0.573)	(0.527)	(0.351)	(0.468)
Constant	No	No	No	No	No	No
Controls	No	Yes	No	Yes	No	Yes
N	540	540	237	237	303	303
Adj. R^2						

Standard errors are clustered across groups and are reported in parentheses. $* p < 0.10$, $** p < 0.05$, $*** p < 0.01$

In Models 3 and 4, we restrict our working sample to observations featuring subjects who contribute less than the group average in the previous period and increase their contributions in the current period. In these models, the number of punishment points received from other group members in the previous period, the average contribution of other group members in the previous period, the previous period's contribution, and the period variable all have a significant effect on the next period's contributions (Model 3). The significance of these variables (except the period variable) is robust to additional controls (Model 4).

In Models 5 and 6, we restrict our working sample to observations featuring subjects who contribute more than or equal to the group average in the previous period and reduce their contributions in the current period. The number of punishment points received from other group members in the previous period, the average contribution of other group members in the previous period, and the previous period's contribution all have a significant effect on the next period's contributions (Model 5). As expected, received punishment points have a negative effect on contributions since they constitute antisocial punishment in this setting. These effects remain significant when controlling for other factors (Model 6).

Punishment decisions

Tables 3 and 4 show that punishment affects contribution decisions. We next investigate the frequency and determinants of punishment decisions.

Table 5 reports punishment frequencies and mean punishment points conditional on the relative contributions of senders and receivers. While punishment

Table 4 Determinants of contributions (OLS)

	(1)	(2)	(3)	(4)	(5)	(6)
Received Points in $t-1$	0.244**	0.191*	0.399***	0.324**	−0.513**	−0.363**
	(0.089)	(0.108)	(0.107)	(0.119)	(0.187)	(0.153)
Other's average contribution in $t-1$	0.612***	0.608***	0.473***	0.426***	0.495***	0.565***
	(0.067)	(0.067)	(0.090)	(0.107)	(0.159)	(0.162)
Contribution in $t-1$	0.336***	0.256***	0.511***	0.493***	0.412***	0.341**
	(0.073)	(0.073)	(0.097)	(0.099)	(0.118)	(0.125)
Period	0.039	−0.031	0.151*	−0.028	−0.107	−0.111
	(0.061)	(0.125)	(0.082)	(0.124)	(0.069)	(0.207)
Final Period	−0.527	−0.262	−0.649	0.007	0.471	0.198
	(0.603)	(0.768)	(0.528)	(0.633)	(1.228)	(1.601)
Constant	No	No	No	No	No	No
Controls	No	Yes	No	Yes	No	Yes
N	361	361	161	161	121	121
Adj. R^2	0.895	0.901	0.956	0.961	0.880	0.898

Standard errors are clustered across groups and are reported in parentheses. $* \, p < 0.10$, $** \, p < 0.05$, $*** \, p < 0.01$

Table 5 Social versus antisocial punishment

Sender's contribution	Punishment			
	Non-zero	Zero	Mean	Median
Higher	338 (52.3%)	308 (47.7%)	1.09	1
Not Higher	204 (17.7%)	950 (82.3%)	0.33	0

The table reports the frequency of social and antisocial punishments along with the mean and median statistics

expenditures are often low, subjects punish those who contributed less than them around 52% of the time. Antisocial punishment, on the other hand, is not prevalent and occurs only in around 18% of cases when a subject observes that a group member contributed at least as much as themselves.[15]

We next disentangle the effects of different variables on the likelihood of punishment. Because punishment expenditures are often low and zero-punishment is commonly observed, we use a series of probit models instead of linear regressions. Our results are reported in Table 6.

Models 1 and 3 use the first-period observations in the P-experiment, while Models 2 and 4 use the same observations with the addition of attitudinal and demographic control variables. When we restrict our sample to observations in which the receiver's contribution is less than the sender's, the two strongest determinants of

[15] Note that the mean punishment expenditures are similar to the values obtained in Herrmann et al. (2008) for Istanbul.

Table 6 Punishment decisions

	(1)	(2)	(3)	(4)	(5)	(6)
Receiver's contribution	−0.100***	−0.151***	−0.085***	−0.018	−0.146***	−0.156***
	(0.034)	(0.051)	(0.030)	(0.035)	(0.021)	(0.023)
Sender's contribution	0.007	−0.048	0.026	−0.023	0.033*	0.025
	(0.032)	(0.058)	(0.041)	(0.050)	(0.020)	(0.024)
Avg. contr. of remaining members	0.067**	0.102**	−0.015	0.078	0.094***	0.106***
	(0.031)	(0.042)	(0.027)	(0.049)	(0.021)	(0.020)
Received sanction points in $t-1$					0.045**	0.045*
					(0.022)	(0.023)
Period					−0.082***	−0.063**
					(0.028)	(0.029)
Final period					0.184	0.082
					(0.161)	(0.188)
Controls	No	Yes	No	Yes	No	No
N	80	80	100	100	1620	1620

Standard errors are clustered across groups and are reported in parentheses. $* \, p < 0.10$, $** \, p < 0.05$, $*** \, p < 0.01$

social punishment in the first period are the target subject's contribution (i.e., the receiver's contribution) and the average contribution of the two remaining group members (Model 1). The significance of these variables is robust to additional controls (Model 2).

Models 3 and 4 consider instances where the sender's contribution is less than or equal to the receiver's in the first period. We find that only the contribution of the target subject (i.e., receiver) significantly affects antisocial punishment in the first period. However, the significance of this variable disappears when we control for other factors.

Models 5 and 6 use all observations in the P-experiment. Model 5 does not include any controls, while Model 6 includes a set of attitudinal and demographic variables. We find that the sender's contribution, the average contribution of the two remaining group members, the punishment points that the subjects received in the previous period, the receiver's contribution, and the period variable all affect the punishment decision (Model 5). All of these variables except the sender's contribution remain significant determinants of punishment decisions after controlling for other factors (Model 6).

Using a linear regression model, Herrmann et al. (2008) find that the amount of social punishment is negatively correlated with the target subject's contribution and positively correlated with the average contribution of the remaining group members in Istanbul (`Table S3A`). In contrast, they find that the amount of antisocial punishment for the same subject pool is negatively correlated with the punisher's contribution and the experiment period (`Table S3B`). In the current study, we obtain the same sign for all the effects, but the effect of the punisher's contribution to antisocial punishment is insignificant. Additionally, Herrmann et al. (2008) do not find an effect of the punishment points received in the previous period on antisocial punishment, while we find a positive effect of this variable in our study.

Details of the agent-based model

In our agent-based model, we follow these steps:

1. We begin our counterfactual analysis by feeding the agent-based model with *actual* first-period contribution data from our 60 subjects in the 15 groups.
2. Subjects either change their contribution or contribute the same amount in the next period. Based on our empirical findings, we link decision rules to subjects' comparison of their last period contribution to the group average: if a subject's last period contribution is less than the group average, they either raise their contribution or keep contributing the same amount in the next period. If their last period contribution is no less than the group average, they either reduce their contribution or contribute the same. In the former case, we use a two-layer contribution rule: in the first layer, we simulate whether a subject retains their contribution based on estimated probabilities defined over received sanction points in the last period (Model 3, Table 3). If the subject does not retain their contribution, in the second layer, we impose that their contribution follows $c_{i,t} = 0.511 \times c_{i,t-1} + 0.473 \times \overline{c}_{-i,t-1} + 0.399 \times \sum_{j\neq i}^{4} s_{j,i,t-1} + 0.151 \times t$. Similarly, if a subject's last period contribution is greater than or equal to the group average, we simulate whether a subject retains their contribution based on estimated probabilities defined over the average contribution of other members in the group during the previous period $\overline{c}_{-i,t-1}$ and last period contribution $c_{i,t-1}$ (Model 5, Table 3). If the subject does not retain their contribution, in the second layer, we impose that their contribution follows $c_{i,t} = 0.412 \times c_{i,t-1} + 0.495 \times \overline{c}_{-i,t-1} - 0.513 \times \sum_{j\neq i}^{4} s_{j,i,t-1}$.
3. We observe that subjects punish others either by a *single* sanction point or they do not punish others in 85.06% of all possible cases. We see that subject i's sanctioning decision on subject j, $s_{i,j,t}$ depends significantly on subject i's received sanction points in the last period $s_{i,t-1}$, the average contribution of the remaining members in the respective group in the current period $\overline{c}_{-i-j,t}$, and the round of the current period t, as well as their and subject j's contribution in the current period: $c_{i,t}$ and $c_{j,t}$. Based on the estimated sanction probabilities from the data, we simulate whether subject i assigns subject j a sanction point (of unity) or not (Model 5, Table 6).[16]

[16] We do not have a history of received sanction points in the first period, so we proceed differently to determine the first-period sanction points. In the first period, we see that subject i's sanctioning decision on subject j, $s_{i,j,1}$ significantly depends on the average contribution of the remaining members in the respective group $\overline{c}_{-i-j,1}$ and j's contribution $c_{j,1}$ for social punishment case (Model 1, Table 6), whereas this decision only depends on j's contribution $c_{j,1}$ for anti-social punishment case (Model 3, Table 6). Based on estimated sanction probabilities from the data, we simulate whether subject i assigns subject j a sanction point (of unity) or not.

4. We move on to the next period and recursively conduct the same steps over periods.

Appendix figures

See Figs. 7, 8.

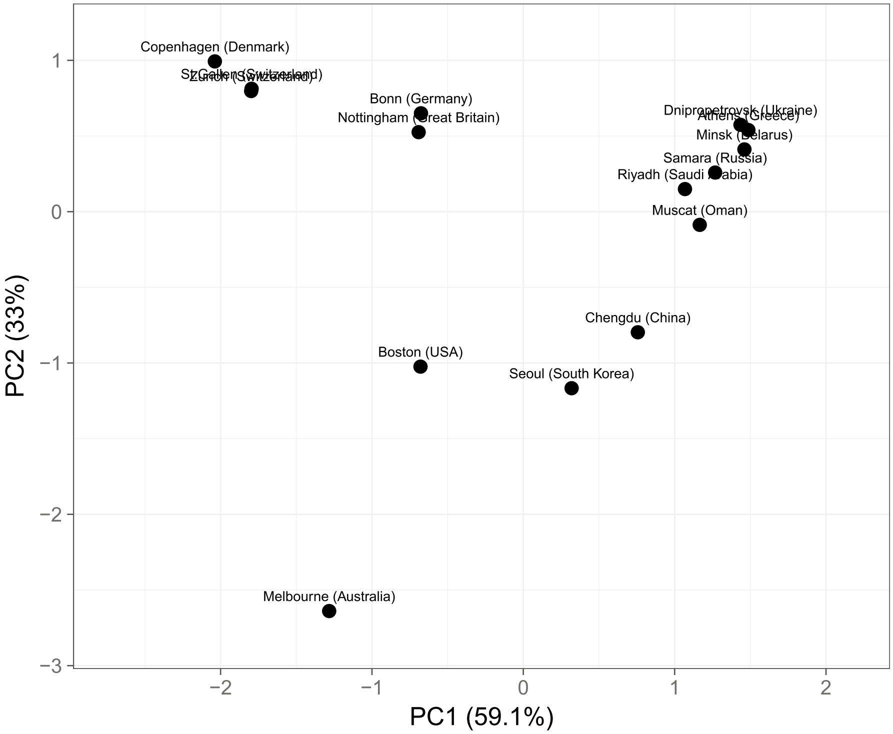

Fig. 7 Principle component analysis (Herrmann et al., 2008). *Notes:* Fig. 7 shows the principle component analysis scores of the fifteen cities by Herrmann et al., 2008 based on (i) straight-line physical distance to Istanbul, (ii) GDP per capita (in 2017 current US dollars), (iii) cultural and psychological distance to Istanbul (Turkey) measured using Muthukrishna (2018)'s `WEIRD` scale index

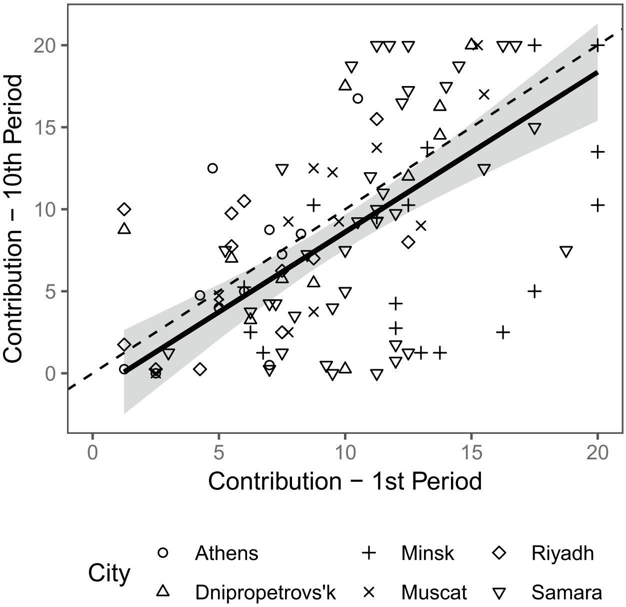

Fig. 8 First-period and last-period contributions by city (FE) (Herrmann et al., 2008).*Notes:* The horizontal axis represents the average contribution in the first period, and the vertical axis represents the average contribution in the tenth period for six cities similar to Istanbul based on the PCA analysis. Each symbol in the figure represents the average contribution of a group. Points below the dashed 45° line indicate that those groups failed to improve upon their average contribution in the first period, while points above the line indicate groups that improved their average contribution in the final period. The solid line represents the linear best-fit line, and the shaded gray areas show the 95% confidence intervals

Data availability All data and replication files can be accessed at https://doi.org/10.7910/DVN/VKXGNG.

References

Andreoni, J., & Miller, J. (2002). Giving according to GARP: An experimental test of the Cconsistency of preferences for altruism. *Econometrica, 70*, 737–753.

Brekke, K. A., Hauge, K. E., Lind, J. T., & Nyborg, K. (2011). Playing with the good guys. A public good game with endogenous group formation. *Journal of Public Economics, 95*(9–10), 1111–1118.

Buchan, N., Brewer, M., Grimalda, G., Wilson, R., Fatas, E., & Foddy, M. (2011). Global social identity and global cooperation. *Psychological Science, 22*, 821–828.

Burlando, R., & Guala, F. (2005). Heterogeneous agents in public goods experiment. *Experimental Economics, 8*(1), 35–54.

Cárdenas, J. C., Dreber, A., Von Essen, E., & Ranehill, E. (2012). Gender differences in competitiveness and risk taking: Comparing children in Colombia and Sweden. *Journal of Economic Behavior & Organization, 83*(1), 11–23.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Erlbaum.

Fehr, E., and Gächter, S. (2000). "Cooperation and Punishment in Public Goods Experiments". *American Economic Review*, 980–994.

Fehr, E., and Gächter, S. (2002). "Altruistic Punishment in Humans". *Nature*, 415, 137-?140.

Gächter, S., & Herrmann, B. (2009). Reciprocity, culture and human cooperation: Previous insights and a new cross-cultural experiment. *Philosophical Transactions of the Royal Society B: Biological Sciences, 364*(1518), 791–806.

Gächter, S., Herrmann, B., & Thöni, C. (2010). Culture and Cooperation. *Philosophical Transactions of the Royal Society of London B: Biological Sciences, 365*(1553), 2651–2661.

Gächter, S., & Thöni, C. (2005). Social learning and voluntary cooperation among likeminded people. *Journal of the European Economic Association, 3*(2–3), 303–314.

Gunnthorsdottir, A., Houser, D., McCabe, K., & McCabe, K. (2007). Disposition, history and contributions in public goods experiments. *Journal of Economic Behavior & Organization, 62*(2), 304–315.

Gunnthorsdottir, A., Vragov, R., Seifert, S., & McCabe, K. (2010). Near-efficient equilibria in contribution-based competitive grouping. *Journal of Public Economics, 94*(11–12), 987–994.

Herrmann, B., Thöni, C., & Gächter, S. (2008). Antisocial punishment across societies. *Science, 319*(5868), 1362–1367.

Isaac, R. M., & Walker, J. M. (1988). Group size effects in public goods provision: The voluntary contributions mechanism. *Quarterly Journal of Economics, 103*(1), 179–199.

Lamba, S., & Mace, R. (2011). Demography and ecology drive variation in cooperation across human populations. *Proceedings of the National Academy of Sciences, 108*(35), 14426–14430.

Muthukrishna, M., Bell, A. V., Henrich, J., Curtin, C. M., Gedranovich, A., McInerney, J., & Thue, B. (2018). "Beyond WEIRD Psychology: Measuring and Mapping Scales of Cultural and Psychological Distance." SSRN Working Paper No: 3259613.

Ones, U., & Putterman, L. (2007). The ecology of collective action: A public goods and sanctions experiment with controlled group formation. *Journal of Economic Behavior and Organization, 62*(4), 495–521.

Ostrom, E., Walker, J., & Gardner, R. (1992). Covenants with and without a Sword: Self-Governance is possible. *American Political Science Review, 86*(02), 404–417.

Sylwester, K., Herrmann, B., & Bryson, J. J. (2013). Homo homini lupus? Explaining antisocial punishment. *Journal of Neuroscience, Psychology, and Economics, 6*(3), 167.

Yamagishi, T. (1986). The provision of a sanctioning system as a public good. *Journal of Personality and Social Psychology, 51*(1), 110.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.