From disapproval to social exclusion: the endogenous formation of non-financial incentives for collectively beneficial behaviours*

Daniel Martinez-Felip[1], Steven G.M. Schilizzi[1], Chi Nguyen[2] and David J. Pannell[1]

[1]Agricultural and Resource Economics, UWA, School of Agriculture and Environment, The University of Western Australia, Crawley, Western Australia, Australia.

[2]Faculty of Economics, University of Economics and Law, Vietnam National University HCM, Vietnam

Abstract

In analysing potential policy responses to improve outcomes in collective-action problems, economists often focus on financial disincentives to reduce the expected gains from free-riding and thereby promote within-group cooperation. In this study, we investigate the potential for groups to develop non-financial disincentives to free riding, thereby promoting convergence towards collectively beneficial actions. Using a within-subjects laboratory experiment, participants play two multi-period public-goods games sequentially: without and then with non-financial incentives activated by allowing for the endogenous formation of a social exclusion mechanism. This is operationalised by allowing participants, at a personal cost, to assign exclusion tickets to group members after observing their contributions: the member(s) having accumulated the most in their group gets excluded from a group activity not involving monetary payoffs nor linked to the main game. First, the threat of receiving exclusion tickets, then the threat of being excluded, and finally actually being excluded work as non-financial social disincentives to free ride. Results show that group members who contribute relatively less receive more exclusion tickets. By imposing expected social costs on relatively low contributors, exclusion or the threat of exclusion enables groups to operate with higher contribution levels, thereby reversing the collective decline in contributions observed in the Baseline public good game. Exclusion is experienced by individuals who consistently contribute less than other group members, and this experience amplifies the effectiveness of the subsequent exclusion threat. Willingness to incur personal costs to enhance the exclusion threat increases over time and it is shaped by more cooperative normative expectations. This effect is particularly pronounced among individuals who perceive norms as tight, especially when higher contributions become more dispersed. In the absence of financial disincentives, these patterns

*Correspondence. Email: martinezfelipd@gmail.com

show how non-financial incentives, shaped by more cooperative normative expectations, can foster group coordination and higher public-good contributions.

Keywords: Non-financial incentives; Social dilemmas; Public goods game; Social exclusion; Laboratory experiments

1. Introduction

Free-riding incentives in collective action problems arise when individuals face a conflict between personal and collective interests, and expect greater personal benefits—both financial and non-financial—from acting selfishly. A robust and well-established finding from controlled laboratory experiments using the Public Goods (PG) game—an experimental game commonly used to study individual decision making under such conflicts—is that individuals often prioritise short-term personal gains, and this undermines the emergence of long-term cooperation (e.g., Fischbacher & Gächter, 2010; Fischbacher et al., 2001; Ledyard, 1995; Zelmer, 2003). Such self-interested behaviour frequently leads to suboptimal outcomes not only for the group but for society at large. Notable real-world examples include overexploitation of common resources (Hardin, 1968), challenges in public health compliance (Betsch et al., 2017; Hershey et al., 1994), and insufficient action to mitigate climate change (Barrett & Dannenberg, 2016). The common failure of groups and communities to deliver optimal collective outcomes highlights the importance of better understanding what drives the emergence of collectively beneficial behaviours within communities.

The potential of *financial* incentives as a mechanism for communities to deter free riding has been extensively studied, particularly in the form of financial punishments for non-cooperative behaviour (e.g., Balliet et al., 2011; Chaudhuri, 2011; Fehr & Gächter, 2000).In laboratory experiments, *financial* punishments are often operationalised charging individuals a financial cost to express social disapproval, thereby imposing a larger financial cost on another individual. Since the employment of such punishments contradicts simple payoff maximisation, they have often been interpreted as an expression of individuals' willingness to adhere to and enforce normative expectations (Eriksson et al., 2017; Fehr & Fischbacher, 2004; Fehr & Schurtenberger, 2018; Kimbrough & Vostroknutov, 2016; Xiao, 2013).These expectations may differ not only in their content but also in their strength: tighter norms are characterised by lower tolerance for deviance, where even small deviations are sanctioned, while looser norms imply greater tolerance (Dimant, 2023; Gelfand et al., 2017; Gelfand et al., 2011).

Yet we still lack a clear understanding of when and why social disapproval escalates within groups, and whether such dynamics are driven by the communication and enforcement of

normative expectations. It also remains unclear whether behavioural responses to social disapproval are motivated by a desire to comply with normative expectations, are financially motivated, or reflect a combination of both. As a result, *financial* punishments can promote cooperation, but they may also backfire due to counter-punishments or an overall reduction in group payoffs (Dreber et al., 2008; Fehr & Rockenbach, 2003; Herrmann et al., 2008), with their effects often disappearing once *financial* incentives are removed (e.g., Brandts & Cooper, 2006; Hamman et al., 2007; Nakagawa et al., 2022).

In contrast to *financial* incentives, real-world communities often rely on purely non-financial social mechanisms to signal normative expectations (Bicchieri, 2006; Ostrom, 1990, 2000; Ullmann-Margalit, 1977). While previous studies have highlighted the potential of social disapproval itself to influence cooperative behaviour (e.g., Brook & Servátka, 2016; Dugar, 2010; Gächter & Fehr, 1999; López-Pérez & Vorsatz, 2010; Masclet et al., 2003; Peeters & Vorsatz, 2013), some have argued that the long-term effect of social disapproval should be contingent to tangible consequences (e.g., Sparks et al., 2024). However, the link between the effectiveness of social disapproval and the social consequences that arise from its accumulation is relatively understudied. This study investigates an example of such a social consequence: (the threat of) social exclusion from community interactions (Francis, 1985; Gruter & Masters, 1986; Kurzban & Leary, 2001; Ostrom, 1990; Ouwerkerk et al., 2005; Williams, 2001).

We aim to extend the literature on *non-financial* incentives in PG games by operationalising social exclusion as an endogenous social consequence derived from accumulated social disapproval within a group. This mechanism aims to reflect the time needed for such a social consequence to be imposed on a shared agreement, which is a fundamental component of group normative expectations enforcement (Bicchieri, 2006).

By isolating non-financial incentives, we create the conditions necessary to address the following questions: (1) Is the *threat of exclusion* capable of reversing a decline in PG game contributions while simultaneously promoting within group coordination at higher levels? (2) Does the *experience of exclusion* enhance or decrease the effectiveness of the threat in fostering contributions and coordination? (3) What drives individuals' willingness to express social disapproval by strengthening the threat of social exclusion? A deeper understanding of such social mechanisms is of importance not only for theoretical understandings but also for policymakers seeking to design cost-effective, community-based interventions that promote sustained cooperative behaviour.

To investigate these research questions, we employ a within-subjects laboratory experiment and place participants in two sequential social contexts: (1) a Baseline repeated PG game, and (2) a repeated PG game where non-financial incentives are activated by allowing for the endogenous formation of a social exclusion mechanism. The formation of this mechanism

involves two distinct but interrelated costs: a personal financial cost incurred when enhancing the threat of exclusion, reflecting a willingness to express social disapproval, and a (expected) social cost associated with the (threat of) social exclusion. In our design, exclusion entails not being able to participate in a non-financial group activity called the Word Formation Game (WFG).[1] We elicited individuals' normative expectations pre- and post-social context with the Krupka and Weber (2013) norm-elicitation method, which has been shown to be robust in eliciting normative expectations (Aycinena et al., 2024; D'Adda et al., 2016; Erkut, 2020; Fallucchi & Nosenzo, 2022).

Anticipating our results, we experimentally show that non-financial incentives that arise within groups can serve as a strong disincentive to free-riding. Specifically, the threat of social exclusion reversed a decline in contributions as rounds of the Baseline PG game proceed. This threat was predominantly directed at those contributing relatively less, imposing an expected non-financial social cost that enabled groups to coordinate on higher contribution levels. Moreover, experiencing exclusion did not backfire or undermine coordination; instead, it amplified and reinforced the effectiveness of the threat. Willingness to incur personal costs to enhance the threat of exclusion increased over time and was shaped by cooperative normative expectations. This effect was particularly evident among tight-norm individuals when higher contributions became more dispersed, reflecting reduced within-group coordination. The remainder of the paper is structured as follows. Section 2 provides an overview of the previous literature on psychological effects of being socially excluded, social exclusion mechanisms in repeated PG games, and highlights this paper's contributions against this literature. Section 3 describes our experimental design. Section 4 presents the results. Section 5 discusses these results and Section 6 concludes.

2. Related experimental literature on social exclusion

Social exclusion has been found to trigger a range of behavioural responses, from prosocial and adaptive behaviours to antisocial reactions (Bernstein & Claypool, 2012; Williams, 2007).[2] These psychological reactions however are not limited to real-life interactions. Zadro et al. (2004) found that even virtual exclusion can trigger psychological responses that are comparable to real-life exclusion.

Within experimental economics, the role of social exclusion has been explored primarily as a tool to solve collective action problems, particularly in the public goods (PG) game. In this context, exclusion is typically conceptualised as a form of monetary punishment. A substan-

[1] More details on how this group activity works are in Subsection 3.2. See Appendix D. for supplementary materials on the experimental instructions.

[2] In psychology, social exclusion, ostracism, and rejection are distinct concepts (Williams, 2007). In this study, we use the term "social exclusion" to encompass all three.

tial body of research in Western Cultures demonstrates that the threat of financial exclusion is a potent mechanism for promoting cooperation. Studies have shown that allowing group members to vote to exclude low contributors—whether by a single vote (Masclet, 2003), a majority rule (Cinyabuguma et al., 2005; Maier-Rigaud et al., 2010), or within an institution that groups explicitly choose to adopt (Dannenberg et al., 2020)—significantly increases contribution levels and effectively targets free-riders. Although these experimental studies were not designed to assess the effectiveness of non-financial incentives in aligning individual and collective interests, they still offer valuable insights into how social exclusion mechanisms operate within groups and their potential to foster cooperative behaviour.

To the best of our knowledge, Davis and Johnson (2015) is the only study to examine the effect of social exclusion mechanism that was not directly tied to a financial incentives on individual behaviour within a group. They experimentally studied a social context where group members could not prevent group members from enjoying the benefits derived from contributions to the public good but had the option to collectively decide who was excluded from a group chat.[3] They found that this form of exclusion effectively deterred free riding, but only after participants had prior experience in a setting without the possibility of exclusion. Importantly, while the chat was prohibited from including direct discussion about the PG game, the authors note that their design encouraged tacit coordination on future contributions. As a result, this setup may have given non-excluded individuals a financial advantage in the game, making it difficult to determine whether increased contributions were driven by the desire to avoid exclusion for non-financial reasons or by the associated financial incentives.

Our study is designed to isolate the impact of purely non-financial incentives in a collective action problem. We introduce a novel experimental design where we extend the literature on non-financial punishments (i.e., social disapproval) by incorporating the threat of exclusion from a group activity that provides no monetary benefit and no strategic advantage in the PG game. We operationalise exclusion as an endogenous social consequence derived from accumulated social disapproval within the group. Thus, our social exclusion mechanism allows us to investigate whether the effectiveness of the accumulated threat of a social consequence (i.e., exclusion) in sustaining contribution behaviour depends on individuals actually experiencing it. This is a question that has remained largely unexplored.

3. Experimental design

This article reports a subset of the data collected in a four-stage, within-subject experiment preregistered on the Open Science Framework (OSF).[4] Figure 1 presents an overview of the

[3] A majority vote rule was needed for a participant to be able to regain access to the group chat in the future.
[4] Registration DOI: https://doi.org/10.17605/OSF.IO/4USJP.

experimental design.

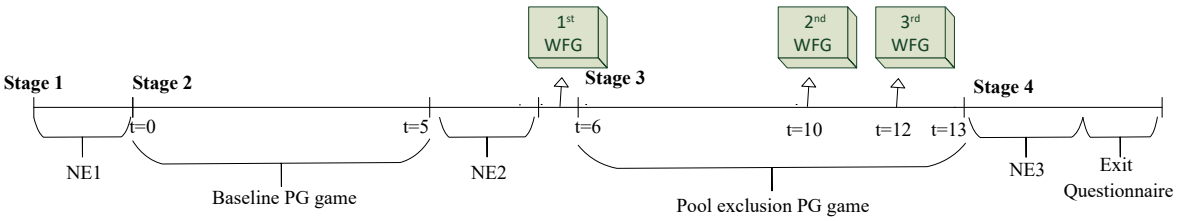

Figure 1: Figure 1: Experiment timeline. NE = Normative Expectations task; WFG = Word Formation Game

To address the research questions of this study, we analyse data from the Baseline PG game and Normative Expectations task 2 (Stage 2) and the Pool Exclusion PG game (Stage 3) of the main experiment. As shown in Figure 1, alongside participants' actual decisions in the PG games, we elicit the evolution of their normative expectations using the Krupka and Weber (2013) method. Specifically, in each session, all participants are presented with four hypothetical scenarios in which three group members are said to have contributed a specific amount to a common group account in the previous round. This amount can take one of four values—0, 350, 650, or 1000—corresponding to 0%, 35%, 65%, and 100% of their endowment, which we refer to as HypS0, HypS35, HypS65, and HypS100. Participants are incentivised to predict the modal social appropriateness of different contribution intervals within each scenario. They learn whether their predictions match the session's modal response only at the end of the experiment. This design feature ensures that the normative expectation tasks do not influence behaviour in the PG games, as participants cannot infer the social appropriateness of different contribution levels from these tasks.

3.1. Baseline PG game

Participants are randomly allocated to fixed groups of four and took part in an incentivised PG game played over 5 consecutive rounds. To avoid end-game effects (Selten & Stoecker, 1986), they are not told how many rounds each part of the experiment will last [5]

At the start of each round, each participant receives an endowment of 50 points. Any contribution to the group account is multiplied by an enhancement factor of $k = 1.6$, and the resulting total is divided equally among all $n = 4$ group members. This produces a marginal per capita return (MPCR) of 0.4 (their share of the 1.6 enhancement factor). In game-theoretic terms,

[5]This approach was chosen so as to collect more robust and representative data across rounds compared to using only one payment-relevant round. Limiting the payment relevance to two rounds, rather than more, also helps to minimize potential wealth and portfolio effects Charness, G., Gneezy, U., & Halladay, B. (2016). Experimental methods: Pay one or pay all. *Journal of Economic Behavior & Organization*, *131*, 141-150. https://doi.org/10.1016/j.jebo.2016.08.010 .

this MPCR creates a conflict between personal and collective interests where the unique subgame perfect Nash equilibrium is zero contribution, while the social optimum requires contributing the entire endowment. [6]

Payoffs for participant i in any round are determined by:

$$\pi_i = 50 - x_i + 0.4 \sum_{j=1}^{n} x_j \tag{1}$$

Where 50 denotes the round's endowment and x_i is i's voluntary level of wealth contributed to the group account.

After each round, participants observe the percentage of their endowment contributed by each group member, identified by round-specific codes to preserve anonymity and prevent tracking of individual contribution across rounds. They also see the total group contribution, their individual share from the group account, and their total payoff for the round.

3.2. Word Formation Game

The social exclusion mechanism is implemented through the (threat of) exclusion from a Word Formation Game (WFG) played with one's group members before re-entering the main PG game. Each WFG presents groups with a unique 5×5 grid of randomly generated letters (see Supplementary Material for an example, Appendix D). Over a two-minute period, group members attempt to create as many words as possible containing at least three letters. Letters can be chosen from anywhere on the board, without any constraint on their configuration (such as having to be on a diagonal), but the same letter cannot be reused within a single word.

At the end of two minutes, participants are told their position in the group ranking based on the number of words they submit. To prevent the rankings from affecting incentives to participate in later WFGs, the "actual" rankings are not revealed. Instead, participants receive a common message and are informed that additional WFGs will take place later in the session, providing further chances to improve their performance. [7]

We introduce the first WFG immediately before the Pool Exclusion PG game to give participants a tangible incentive to avoid exclusion and to make clear what they can lose if excluded from subsequent WFGs.

[6] A conflict between personal and collective interests arises when $< \text{MPCR} < 1 < (\text{MPCR} \times n)$.

[7] To control for participants' motivation to improve their ranking in later analyses, particularly the value they place on this group activity, we asked them in the exit questionnaire how much they enjoyed the WFG. On average, they reported an enjoyment score of 7.45 out of 10 (see Supplementary Material, Appendix D).

3.3. Pool Exclusion PG game

In the Pool Exclusion PG game, participants interact with the same group members for 8 rounds of the PG game, increasing the total number of rounds each participant played with their group members to 13. In addition, two additional WFGs take place at the end of rounds 10 and 12. In the Pool Exclusion PG game, group members can express social punishment to express a normative signal: at a personal cost, they can allocate exclusion tickets to group members, thereby enhancing their threat of exclusion from the upcoming WFGs. Assigning an exclusion ticket costs 1 point, [8] deducted from the player's payoff for that round, as shown in Eq. (2):

$$\pi_i = 50 - x_i + 0.4 \sum_{j=1}^{n} x_j - ct_i \tag{2}$$

50 denotes the round's endowment, x_i is player i's voluntary contribution to the group account, t_i is the number of exclusion tickets assigned by player i, and c is the cost of assigning each exclusion ticket. To limit strategic play, participants were not told in advance when the additional WFGs would occur. At each WFG, the participant with the most tickets was excluded; if two group members have the same highest number of tickets, then they are both excluded. This made exclusion decisions both collective and endogenous, characteristic of pool exclusion. Unlike the WFGs in Charness et al. (2014), which served the purpose of team-building exercises, our WFGs acted as a non-financial incentive embedded within group interactions, introducing the possibility of social consequences through cumulative social disapproval.

3.4. Exit Questionnaire

At the end of the experiment, all participants complete a short exit questionnaire collecting data on gender, age, environmental attitudes, demographics, perceived risk preferences regarding the threat of exclusion, the number of times they were excluded, and their reactions to actual or threat of exclusion (see Supplementary Material, Appendix D).[9] Participants' accumulated points were then converted to Australian dollars (AUD) at a rate of AUD 0.08 per point and paid in cash.

[8] We set a 1:1 ratio for the cost of an exclusion ticket for two reasons: Firstly, the cost associated with assigning exclusion tickets is designed to reflect the willingness of individuals to punish other group members based on their contribution levels. Secondly, a moderate cost enables participants to assign different levels of punishment based on others' contributions, thereby allowing them to express graduated levels of social disapproval.

[9] Note that self-reported assessments of one's own risk attitudes have been found to be a powerful predictor of one's risky behaviour Dohmen, T., Huffman, D., Schupp, J., Falk, A., Sunde, U., & Wagner, G. G. (2011). Individual risk attitudes: Measurement, determinants, and behavioral consequences. *Journal of the European Economic Association*, *9*(3), 522-550. https://doi.org/10.1111/j.1542-4774.2011.01015.x

3.5. Procedures

The preregistered experimental design was approved by the UWA Human Research Ethics Office committee (2023/ET000980).

The experiment was programmed and run using the experimental software oTree (Chen et al., 2016). and conducted in the Behavioural Economics Laboratory at The University of Western Australia (BELab). At the start of each session, instructions were explained in detail, after which participants answered comprehension questions to ensure understanding. Prior to the data collection, pilot sessions were conducted with UWA students to: (1) verify that the instructions were clear and easily understood, (2) assess the functionality of the social exclusion mechanism in the laboratory setting, and (3) generate preliminary data for power analysis.

Participants (N = 100) were recruited via the online system Sona Systems.[10] We followed the standard economics procedure where, in our case, undergraduate and postgraduate students from various disciplines were invited to participate in the experiment. The experiment was run over seven sessions, with a total of 25 groups of four participants each. While the number of participants per session varied, all groups consisted of four members. Contributions in the PG games were incentivised and determined at the group level. Average participant earnings were 27 AUD. The order of treatments was intentionally not reversed. We assumed that participants needed to first experience behaviour in the Baseline PG game to form group-relevant expectations, which could then be enforced in the Pool Exclusion PG game. Experiencing the Baseline PG game first was therefore a necessary condition for the normative signals in the Pool Exclusion PG game to be meaningful.

3.6. Preregistered hypotheses

Before running the experimental sessions, we preregistered the following hypotheses. We anticipated that non-financial incentives, activated through the endogenous social exclusion mechanism, would have effects on contributions similar to those observed in PG games where exclusion functioned as a form of financial punishment (see Section 2). In particular, we hypothesise that the threat of social exclusion from a non-financial group activity can reverse a declining trend in contribution levels **(H1a)**. Additionally, we expect that an increase in the accumulated exclusion tickets received by an individual leads to higher contribution levels in the following rounds **(H1b)**, and that, on average, excluded individuals will increase their

[10] To determine the sample size for our experiment, we conducted prior power analyses for all preregistered hypotheses using the G*Power software. We accounted for matched-pairs data, with the Wilcoxon signed-rank test serving as the reference for within-subject comparisons, for which we assumed a medium effect size of 0.5 standard deviations. For regression-based analyses, we assumed a medium effect size of $f^2 = 0.15$. The analyses suggest that a total 100 participants are needed to reach an overall power of 80% at the 5% significance level.

contributions after being excluded, thereby making the experience of exclusion enhance the effectiveness of a subsequent threat **(H1c)**. Regarding willingness to enhance the threat of exclusion, we hypothesise that individuals who consider higher contributions as most socially appropriate will be more likely to assign exclusion tickets to group members contributing less than that expectation **(H2a)**. Over the 8 rounds of the pool exclusion PG game, we also hypothesise a negative or non-relation between the number of exclusion tickets assigned and the number of exclusion tickets received by an individual **(H2b)**.[11] Finally, we anticipate a negative relation between the total number of exclusion tickets assigned at the group level and the group's average contribution levels **(H2c)**.

We analyse panel data to examine how individuals adjust their contribution behaviour after experiencing, or being exposed, to the threat of social exclusion. We also investigate the factors that drive individuals' willingness to reinforce this threat by assigning exclusion tickets.

4. Results

The results section is structured as follows. First, we examine the impact of the threat of social exclusion from the WFGs on contributions in the PG game. Next, we analyse how actual social exclusion influences the effect of this threat on contribution behaviour. Finally, we investigate the factors underlying the emergence and functioning of the endogenous social exclusion mechanism.

4.1. Threat of exclusion on PG contributions

The experimental timeline in Figure 1 illustrates that, rounds 6 to 10 of the Pool Exclusion PG game, participants faced the threat of exclusion from the second WFG. Figure 3 displays how average contributions to the PG evolved, with 95% confidence intervals, comparing periods with and without the exclusion threat. We also examined how group interdependence changed over time by calculating the intraclass correlation coefficient (ICC), defined below, for contributions separately for each round. We fitted a linear mixed-effects model with a random intercept for each group, and calculated the ICC using the standard formula: ICC (%) $= \frac{\sigma^2_{between-group}}{\sigma^2_{between-group} + \sigma^2_{within-group}} \times 100$.

The ICC quantifies the proportion of total variance in contributions attributable to group-level differences (Hox et al., 2017; Lüdecke et al., 2021). Ranging from zero to one, in the context of PG game contributions, an ICC of zero indicates that individual contributions vary independently of group members, while an ICC of one indicates that group members contribute identical amounts. Consequently, computing the ICC over rounds provides an

[11] Considering that participants do not know how many exclusion tickets their group members have accumulated, and that assigning exclusion tickets to potentially avoid exclusion comes at a cost, we hypothesise that self-serving punishment will not be present in our study.

understanding of PG game contributions beyond what average contribution levels alone can reveal, distinguishing between changes driven by individual strategic choices and those driven by group-level dynamics and coordination (Nielsen & Pfattheicher, 2024). This evolution is also shown in Figure 2.

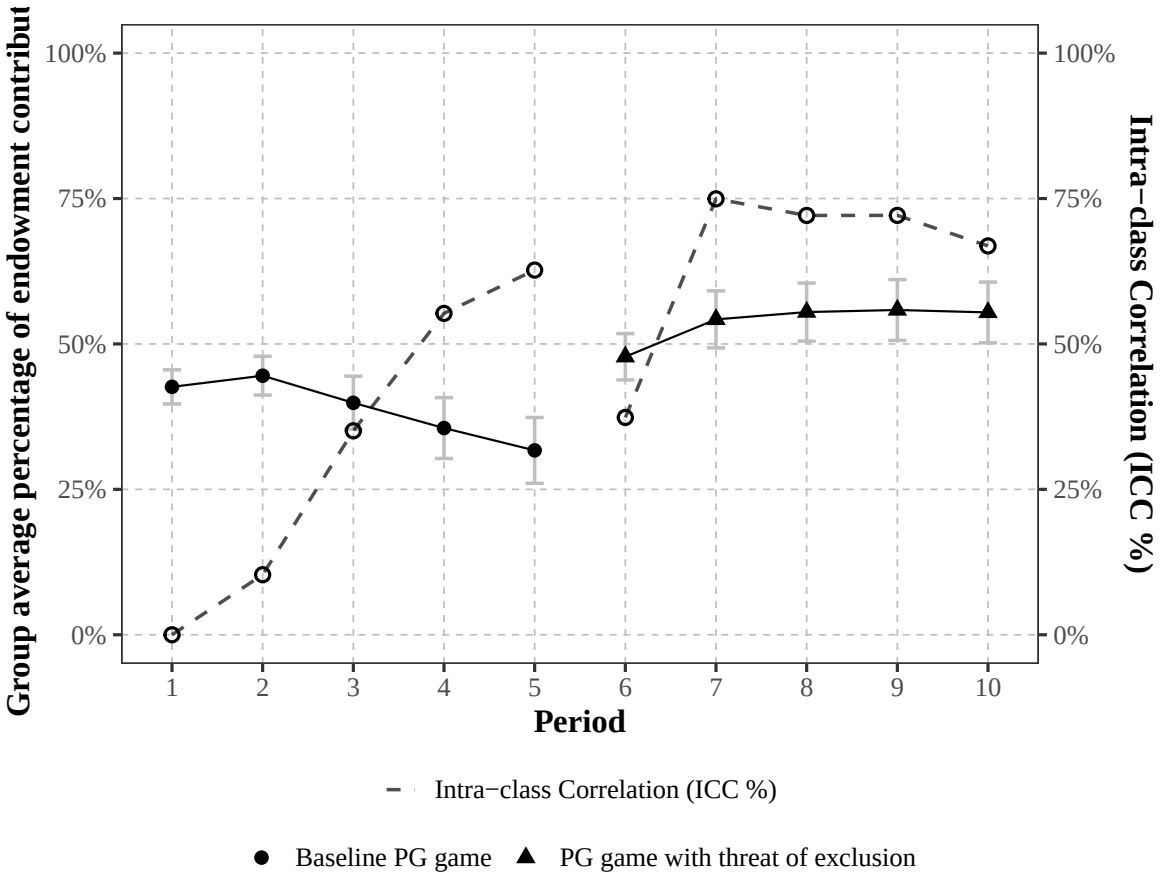

Figure 2: Group average PG contributions per round (percent of endowment) in the Baseline PG game and the PG game with threat of exclusion, with error bars showing ±1 standard error of the mean, and the intra-class correlation (ICC, %).

From Figure 2, it can be seen that contributions exhibited the common decline-over-time pattern in the Baseline PG game. In contrast, during the threat of exclusion PG periods this trend was reversed, with significantly higher average contribution levels observed (see Appendix A for statistical support). Thus, we find **direct support of H1a**. Importantly, previous experimental evidence, allowed us to assume that if the Baseline PG game had extended to ten rounds, contribution levels would have continued to decrease, or at least not increase. This strengthens the case that the positive and highly significant *period* and *Threat of exclusion* interaction coefficient was indeed driven by the presence of the threat of exclusion.

The ICC began at zero in the first period, indicating no initial significant between-group differences, consistent with the absence of prior interaction, or opportunity for adaptation or reciprocation. The ICC then increased over rounds, exceeding 60% by the end of the Baseline PG game. This increase, driven by greater between-group variance and lower within-group variance (see Appendix A), reflects the emergence of group-specific contribution patterns, with members converging toward lower within-group contributions.

Notably, these patterns were disrupted at the start of the Pool Exclusion PG game, with the ICC dropping to 37% in period 6. From period 7 onwards, the ICC rose sharply and remained high ($\approx$72%) when the threat of exclusion was present.[12] This sustained convergence resulted primarily from significantly reduced within-group variability, while between-group variability remained relatively stable (see Appendix B). This dynamic suggests that the increase in average contributions under the threat of exclusion was supported by convergence within the group towards systematically higher contributions.

Importantly, participants faced a heterogeneous threat of exclusion based on the number of exclusion tickets received at time $t-1$. We continue the analysis by examining the effect of receiving exclusion tickets at $t-1$ on contributions at t, that is the marginal increase in accumulated exclusion tickets. Since contributions at $t-1$ are likely to influence contributions at t, it is necessary to account for prior contribution behaviour. However, because exclusion tickets at $t-1$ are directly determined by contributions at $t-1$, including the latter as a control would introduce endogeneity. To avoid this, we instead use the cumulative average of individual contributions up to and including round t–1, which provides a more stable measure for past behaviour while reducing endogeneity concerns. We employ mixed-effects regression with random effects for each participant and we cluster standard errors at the group level. Results are presented in Table 1.

[12] Although not shown in Figure 3, the computed ICC was 78.3% in round 11 and peaked at 79.3% in round 12 before decreasing to 73.7% in round 13, driven by a continued decrease in within-group variance. See Appendix B.

Table 1: Mixed-effects estimates of PG contributions in the Pool exclusion PG game at t: Effect of receiving exclusion tickets at $t-1$.

Dependent variable: i's PG contribution at t	**Coefficient**	**Standard error**	**t statistic**	**p-value**
(Intercept)	8.996	5.051	1.780	0.0970^{*}
i's average PG contribution (Rounds 6 up to and including $t-1$)	+0.815	0.044	18.215	$< 0.001^{***}$
i's exclusion tickets received at $t-1$	+0.668	0.194	3.438	0.0257^{**}
Period	-0.141	0.416	-0.338	0.738
i's exclusion tickets received at $t-1$ × Period	-0.219	0.046	-4.705	0.0087^{***}

Only the first 5 rounds of the Pool Exclusion PG game are considered to focus on the threat of exclusion. Additional controls: gender, demographics, risk, environmental attitudes, age, WFG enjoyment. We did not include both the total number of exclusion tickets accumulated up to round $t-1$ and the marginal number of tickets received in round t in the same model due to multicollinearity concerns (Variance Inflation Factor > 5). Standard errors are clustered at the group level. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

As shown in Table 1, average group contributions had a positive and significant effect on next round contributions. Importantly, on average, receiving exclusion tickets at $(t-1)$ led to an increase in PG contributions in the subsequent round.[13] The semantic analysis of responses to Questions 8 and 9 in the exit questionnaire indicates that 60.71% of participants avoided exclusion by increasing their contribution levels, reinforcing the statistical findings presented. [14] The negative and significant coefficient on the interaction between exclusion tickets and period suggests that, while the threat of exclusion initially increased contributions, it did not lead to continuous increases as hypothesised. Rather, it established higher contribution levels which were then sustained over time. Thus, we find **partial support for H1b.**

In addition, we estimated the same model but using the total number of exclusion tickets accumulated by a participant up at round $t-1$ as a predictor of their contribution in the next round, and it was not statistically significant. These results suggest that the marginal increase in the threat of exclusion was a more salient determinant of a participant's contribution in the following round than the accumulated social sanction history. One possible explanation of this finding is that participants were not informed of the number of exclusion tickets their group members had accumulated at each round. As a result, newly received tickets may have served as a salient and immediate signal of an increased personal threat of exclusion.

[13] The same results are obtained when using the cumulative average excluding round $t-1$, fully addressing endogeneity concerns. However, we present results up to and including round $t-1$ to maximise the number of observations.

[14] Due to data loss, we were only able to obtain qualitative data on reactions to avoiding exclusion from 84% of participants. Consequently, the semantic analysis and reported percentages are based solely on this subset of data.

Finding 1: When the threat of exclusion was present, the standard decline in PG game contributions was reversed, and groups converged at higher contribution levels. However, the effect of exclusion threat on contribution levels was strongest at the beginning, establishing higher contributions that were then sustained.

4.2. Effect of actual exclusion experience on threat effectiveness

We now examine how experiencing actual social exclusion from the WFGs influenced the effectiveness of the threat of exclusion on contribution behaviour in the PG game. Specifically, we compare contribution levels before and after the second and third WFGs (the first one being the WFG with no exclusion possibilities played immediately prior to the Pool Exclusion PG game), distinguishing between excluded and non-excluded participants. Importantly, following the second WFG, all exclusion tickets were reset, and participants continued interacting with the same group members under the same endogenous exclusion rules for the third WFG. This design allows for redemption and enables us to identify and compare four distinct participant types based on their exclusion history: (1) excluded only from the second WFG, (2) excluded only from the third, (3) excluded from both the second and third, and (4) never excluded. Figure 3 presents four panels. Panels A and B depict the evolution of average contributions and deviations from group contribution averages, respectively, before and after the second WFG for excluded (28% of participants) and non-excluded (72%) participants. Panels C and D show the corresponding patterns before and after the third WFG, differentiating between those excluded only from the second WFG (23%), only from the third WFG (20%), from both the second and third WFGs (5% of participants), and those never excluded (52%).

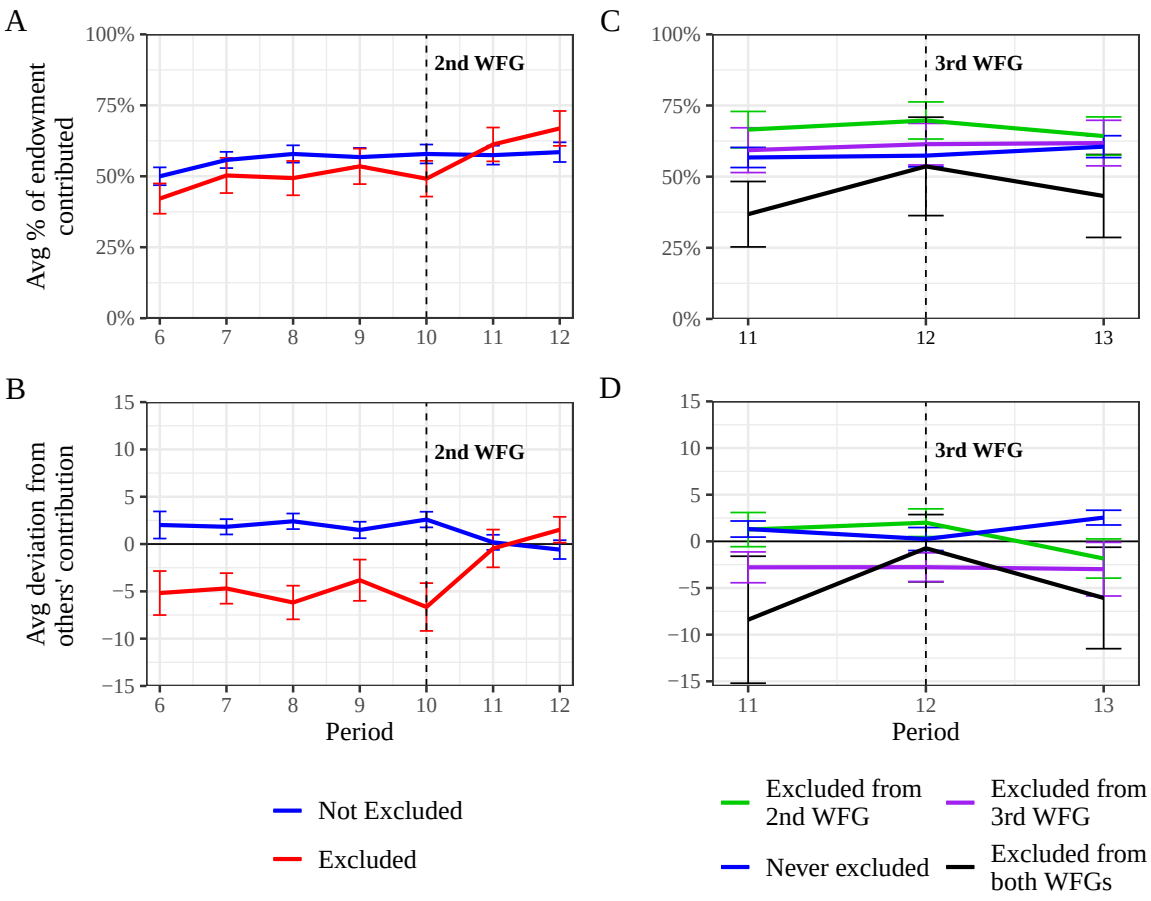

Figure 3: Average PG contributions (in % of endowment) and deviations from others' contributions, both shown with standard errors of the mean, by exclusion category before and after second and third word formation games (WFGs)

Panels in Figure 3 show that, on average, for the 52% of participants who were never excluded, the threat of exclusion helped sustain high contribution levels over time. By contrast, participants who were ultimately excluded, whether from the second or third WFG, consistently contributed less than their group members.

Finding 2: Social exclusion was primarily experienced by those individuals that consistently contributed relatively less than their group members.

However, given that social exclusion was endogenously determined, direct contribution comparisons between excluded and non-excluded participants during the post-exclusion period are not possible, as the observed differences between the two groups do not solely reflect the causal effect of exclusion but also pre-existing differences in their behaviour (which determined receiving exclusion tickets).

To estimate the change in the threat's effectiveness on contributions after experiencing exclusion, we employ a *Difference-in-Differences (DiD)* estimator. This approach relies on the *parallel trends assumption*, that is, in the absence of exclusion, the contribution trajectories of excluded and non-excluded participants would have followed similar trends.

For exclusion from the second WFG, Panels A and B of Figure 3 support the plausibility of this assumption: during the pre-exclusion period (Rounds 6 to 10), where only the threat of exclusion was in place, *excluded* and *non-excluded* participants exhibited closely aligned trends in both average contribution levels and their deviations from others' contributions. In addition, we used the synthetic control method to construct a counterfactual for excluded participants derived from a weighted average of comparable non-excluded individuals. The resulting synthetic control further reinforced the credibility of the *parallel trends* assumption (see Appendix C).

Importantly, to study the effect of exclusion from the third WFG, Panels C and D of Figure 3, participants *Excluded from both WFGs* can only be compared to those *Excluded only from the second WFG,* while participants *Excluded only from the third WFG* can only be compared to those *Never excluded*. We assume parallel trends between these comparison groups based on visual inspection of Panels C and D. However, this cannot be empirically verified due to the limited number of pre-exclusion periods. For the same reason, a synthetic control was not constructed.

We estimated the *DiD* model using a mixed-effects model with individual-level random effects, and standard errors clustered at the group level. Results are presented in Table 2.

Table 2: Mixed-effects Difference-in-Differences (DiD) estimates of PG contributions – Effect of exclusion from the second and third WFGs

Variable / Effect Description	**Estimate**	**SE**	**t-value**	**p-value**
Effect of exclusion from second WFG				
DiD estimate: Excluded (vs Not excluded) $\times$ post-exclusion period	+6.411	1.466	4.371	$< 0.001^{***}$
Effect of exclusion from third WFG				
DiD estimate: Excluded 3rd WFG only (vs. Never excluded) $\times$ post-exclusion period	-1.015	3.365	-0.301	0.767
DiD estimate: Excluded both WFGs (vs. Excluded 2nd WFG only) $\times$ post-exclusion period	0.934	4.777	0.195	0.851

Only DiD estimates of the causal effect of exclusion on contributions are reported in the regression table. Effect of exclusion from 2nd WFG: The pre-exclusion period includes rounds 6 to 10, while the post-exclusion period includes rounds 11 to 12. Effect of exclusion from 3rd WFG: The pre-exclusion period includes rounds 11 to 12, while the post-exclusion period includes round 13. Standard errors are clustered at the group level. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$. Additional controls: gender, demographics, risk, environmental attitudes, age, WFG enjoyment.

As shown in Table 2, we found a positive and highly significant interaction effect between being excluded and the post-exclusion period. This indicates that the experience of exclusion significantly enhanced the subsequent effectiveness of the exclusion threat, leading to a substantial increase in contributions and providing **support for H1c**. Moreover, the exclusion experience from the second WFG also improved the effectiveness of the threat of exclusion in coordinating group behaviours around these increased contributions. The ICC increased from 66.8% to 78.3%, as detailed in Appendix B.

In contrast, when examining the effect of exclusion from the third WFG presented in Table 2, the non-significant DiD effect of both comparison groups suggests that experiencing exclusion at this stage did not lead to further subsequent increases in the effectiveness of the exclusion threat. This may reflect the fact that contribution levels were already relatively high among those not excluded in the second WFG and had already risen among those who were. Notably, individuals excluded from both WFGs (5% of participants) continued contributing less than others, both before and after exclusion, which we use as a criterion to say that only 5% of participants were highly unresponsive to non-financial incentives.

Rather than leading to additional increases, exclusion from the third WFG appeared to reinforce previously the previously increased effectiveness of the exclusion threat on PG contributions. However, it reduced the effectiveness of the threat of exclusion on contribution coordination, as evidenced by a decrease in the ICC from 79.2% to 73.7% (see appendix B). Despite this reduction, group coordination at higher contribution levels following both exclusion events remained relatively higher than it was before the first exclusion event. However, the limited number of post-exclusion periods following the third WFG prevents us from assessing whether exclusion had any delayed effects beyond the immediate round.

Finding 3: Being socially excluded increased and reinforced the effectiveness of the exclusion threat on contribution levels.

4.3. Endogenous emergence of social exclusion as a sanction: assignment of exclusion tickets

We now examine *what* drives the emergence of the social exclusion mechanism.

Emergence of an endogenous sanction against group members seen as contributing too little

The assignment of costly exclusion tickets captures individuals' willingness to incur a cost to express social disapproval by enhancing the threat of social exclusion. Figure 4 provides a visual representation of the endogenous emergence of the social exclusion mechanism. Panel (1) shows the sample average (N = 100) number of exclusion tickets *received* over time by

participants contributing within different ranges. Panel (2) shows the sample average of exclusion tickets *assigned* over time.

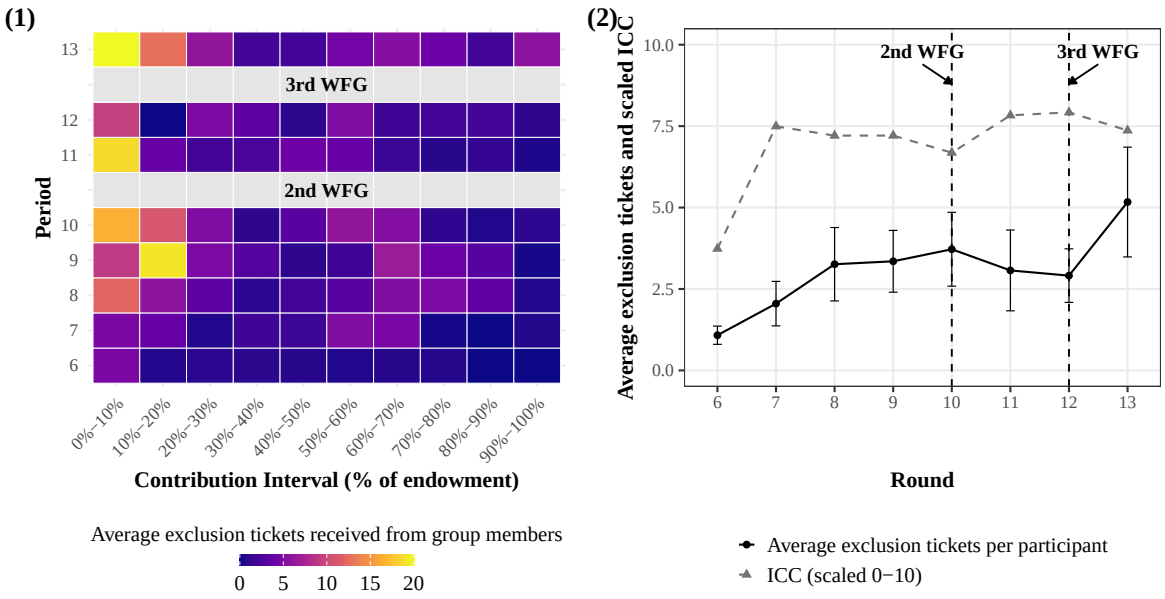

Figure 4: (1) Average exclusion tickets a participant received from group members, by own contribution interval (% of endowment) and period. (2) Average exclusion tickets assigned per participant by period (± SE across groups), with ICC of contributions (ICC rescaled to 0–10 for display). Vertical dashed lines mark the two exclusion events (periods 10 and 12).

As shown in Panel (1), the enhancement of the social exclusion threat was primarily directed at low contributors, and this disapproval increased over rounds. Panel (1) also indicates that in the final round, there was behaviour consistent with antisocial punishment, targeting members who contributed 90% to 100% of their endowment. Panel (2) shows that in the first round of the Pool Exclusion PG game (round 6), participants assigned an average of 1.08 exclusion tickets after observing others' contribution intervals. This number gradually increased over the subsequent rounds where only the threat of exclusion was present, declined in the two rounds following the second WFG, and then increased sharply after the third WFG in the final round to an average of 5.17 tickets. Participants were not informed about the total number of rounds in the experiment and thus may have anticipated the possibility of a third exclusion event. Therefore, the increase in exclusion tickets observed in the final round reflects a greater willingness to enhance the threat of exclusion in reaction to not seeing contribution levels increasing. After the initiation phase from round 6 to round 7, these dynamics were negatively correlated with within-group coordination on contributions (ICC).

Assigning an exclusion ticket incurred a cost of 1 point deducted from the participant's payoff

(Eq. 2). Given average participant payoffs of 64.3 points in round 6 and 68.8 points in round 13 at the time of ticket assignment, the increase in exclusion ticket assignments represents a larger share of payoffs dedicated to enhancing the threat of exclusion: 1.67% in the first pool-exclusion round versus 7.57% in the last.

Finding 4: Exclusion tickets assigned to members contributing between 0% to 20% of the endowment increased substantially over time, rising on average by a factor of 4.1 per participant.

How normative expectations drive willingness to sanction at cost to oneself

We now turn to the drivers of the emergence and intensification of the social exclusion mechanism. Leveraging our individual-level panel data, we examine how participants' perceived social norms shaped their willingness to assign exclusion tickets to group members. To do so, we incorporate data from the normative expectation elicitation task conducted prior to the Pool exclusion PG game (NE2), which captured in each hypothetical scenario both the contribution level perceived as most socially appropriate and its perceived normative strength.

To understand how normative expectations shaped individuals' willingness to assign exclusion tickets at personal cost, one needs to recognise their context-dependent nature. As shown in our parallel study (Martinez-Felip et al., 2025), norm profiles and perceptions of socially appropriate behaviour reflected a tendency to judge as most appropriate contributions similar to others. Accordingly, identifying which normative expectation participants are likely to activate in each round of the Pool Exclusion PG game requires consideration of the actual contribution behaviour they experienced.

To capture this, we aligned the elicited normative expectations with the behaviour participants actually experienced in the game. In each round, the group's average contribution determined which norm scenario was active (HypS0, HypS35, HypS65, or HypS100).[15] For each participant i, we then classified into three groups every group member's contribution relative to i's perceived most appropriate contribution level (PMACL) for the active scenario: as a negative deviator (below i's PMACL), non-deviator (matching i's perceived social norm), or positive deviator (above i's perceived social norm). The PMACL was measured using the coefficient of variation (CV) of appropriateness ratings: participants with a CV lower than the median were classified as perceiving loose norms, and those with a CV higher than the median as perceiving tight norms.

Given that exclusion tickets are count data, we analysed the number of tickets assigned using a negative binomial generalised mixed effects model (GLMM) with random intercepts for in-

[15] The rule was as follows: when the group's average contribution fell between 0% and 17.5% of the endowment, the activated scenario was HypS0; when it was above 17.5% and up to 52.5%, HypS35; when it was above 52.5% and up to 87.5%, HypS65; and when it was 87.5% or higher, HypS100.

dividuals and groups. To ensure robust inference, standard errors were clustered at the group level, accounting for potential within-group correlation beyond that captured by the random effects. The model controls for the round-specific information present when participants made their assignment decisions: the within-round dispersion of contributions (standard deviation), the individual's own contribution in round t, the number of tickets already accumulated up to $t-1$. Since the number of participants in each norm-deviation category varied across rounds (e.g., sometimes more negative than positive deviators or non-deviators), we included the log of the category size as an offset in the model, thereby standardising the estimated effects of the norm-deviation categories by the number of participants in each category. The dependent variable is the count of exclusion tickets that individual i assigned to each category in a given round. [16]

Table 3 reports the incidence rate ratios (IRR) from this model, with 95% confidence intervals and significance levels.

[16] We assessed zero inflation using the DHARMa zero-inflation test, which showed no evidence of excess zeros (ratioObsSim = 1.01, $p = 0.944$). We also found evidence of overdispersion in the Poisson GLMM (Pearson dispersion statistic = 2.70), indicating that the conditional variance exceeded the mean. Accordingly, we estimated a negative binomial GLMM, which provided a better fit and accounted for the overdispersion.

Table 3: Incidence Rate Ratios (IRR) of the Negative Binomial GLMM of exclusion tickets assigned by i to group members at t

Predictor	**IRR**	**95% CI**	**p-value**
SD of group contributions at t	1.03	[0.990, 1.073]	0.140
Own contribution at t	0.98	[0.961, 1.007]	0.192
Period	1.11	[1.044, 1.177]	$< 0.001^{***}$
Exclusion tickets accumulated up to $t-1$	1.01	[1.002, 1.023]	0.020^{**}
Deviation category (ref = Non-deviators)			
Negative deviators at t	3.04	[1.99, 4.462]	$< 0.001^{***}$
Positive deviators at t	0.53	[0.341, 0.850]	0.007^{***}
Norm strength category (ref = Loose)			
Tight norm perception	0.94	[0.498, 1.788]	0.859
Tight norm perception $\times$ SD of group contributions at t	1.09	[1.051, 1.131]	$< 0.001^{***}$

Notes: Deviation categories. Group members were classified relative to participant i's perceived social norm, elicited in the NE2 task. In each round, the group's average contribution determined the active hypothetical scenario (HypS0, HypS35, HypS65, or HypS100), and i's perceived norm for that scenario was used to classify group members as non-deviators (at the norm), negative deviators (below the norm), or positive deviators (above the norm). Norm strength was measured using the coefficient of variation (CV) of appropriateness ratings: participants with lower CVs were classified as perceiving loose norms, and those with higher CVs as perceiving tight norms. The model included controls for gender, demographics, environmental attitudes, age, and WFG enjoyment, with random intercepts for participants ($n = 100$) and groups ($n = 25$). An offset equal to the log of the number of group members per deviation category adjusted for availability of targets. Standard errors were clustered at the group level. IRR = incidence rate ratio; 95% CI = confidence interval. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$. Model fit: AIC = 2873.4, BIC = 2981.6, Log-Lik = -1414.7.

Table 3 shows that normative expectations played a central role in explaining participants' willingness to enhance the threat of exclusion, and showed the strongest effects in the model. The incidence rate ratio for negative deviators was well above 1 and highly significant (IRR = 3.04, p value < 0.001), indicating that, compared to non-deviators, participants assigned nearly three times as many tickets per available member to group members who contributed less than their perceived norm. By contrast, the incidence rate ratio for positive deviators was significantly below 1 (IRR = 0.53, p value = 0.007), showing that participants assigned about 45% fewer tickets per available member to those who contributed more than their perceived norm. These findings support H2a, namely that individuals who perceive higher contributions as most socially appropriate are more likely to sanction group members who contribute less than this expectation.

In addition, the significant interaction between tight-norm participants and the standard deviation of group contributions (IRR = 1.09, p < 0.001) shows that those with tighter normative expectations were especially responsive to contribution variability within the group. In other

words, as higher contributions became more dispersed, tight-norm participants consistently assigned more exclusion tickets than loose-norm participants, consistent with lower tolerance for deviant behaviour. Greater dispersion in contributions therefore increased willingness to sanction, and this effect occurred primarily among participants perceiving a tight norm. This helps explain the negative association between coordination and ticket assignment shown in Panel (2) of Figure 4: as contribution levels became more heterogeneous, those with tighter expectations drove a stronger enforcement response, which can be interpreted as an effort to push the group to all contribute higher levels.

Responses to the exit questionnaire show that a minority of participants (13.1%) reported assigning exclusion tickets mainly to avoid being excluded themselves. This further suggests that participants valued being part of their reference groups activities (WFGs) and perceived exclusion as a meaningful sanction. Consistent with this, Table 3 shows that the effect of accumulated exclusion tickets up to t – 1 was statistically significant; thus, we do not find support for **H2b**. However, this effect was very small (IRR = 1.01, p value = 0.02) compared with the strong and systematic effects of motives linked to norm-deviation ($\approx 3\times$ more tickets assigned to negative deviators, $\approx$45% fewer to positive deviators). This suggests that while the exclusion mechanism activated mixed motives, the dominant driver of punishment was the enforcement of cooperative normative expectations.

In addition, a Pearson product-moment correlation revealed a significant negative relationship at the group level between average contributions and the average number of exclusion tickets assigned ($r = -0.22$, $p < 0.001$). Thus, groups with higher average contributions assigned fewer exclusion tickets. This supports the intuition that in high-contributing groups there was less need for exclusion tickets to enforce contributions, as hypothesised in **H2c**.

Finally, considering the patterns in both panels of Figure 4 together with the results in Table 3, our findings suggest that more cooperative normative expectations emerged in the pool exclusion PG game. The rising number of exclusion tickets assigned over time to group members contributing below the normative expectation reflects the emergence of shared normative expectations—as shown in our complementary study (Martinez-Felip et al., 2025)—and an increasing willingness to enforce them. This can be interpreted as the emergence of a (cooperative) social norm at the group level.

Finding 5: Willingness to incur personal costs to enhance the threat of exclusion was negatively associated with group contribution coordination and was primarily driven by cooperative normative expectations. Participants assigned substantially more exclusion tickets to group members who contributed below their normative expectation than to those who adhered to it, while members who contributed above the norm were sanctioned significantly less.

We summarise our results relative to our initial hypotheses in Table 4.

Table 4: Preregistered hypotheses and corresponding results

Contributions to the PG	
H1a: Contribution levels are significantly higher when the threat of exclusion is present.	**Supported**
H1b: An increase in the aggregated exclusion tickets received by an individual increases his/her contribution levels in the following round.	**Partially supported.** Receiving exclusion tickets led to an increase in PG contributions in the following round. However, this effect weakened over time.
H1c: On average, excluded individuals will contribute more after being excluded, thereby making the experience of exclusion enhance the effectiveness of a subsequent threat.	**Partially supported.** Exclusion increased the effectiveness of the subsequent threat. After being excluded from the second WFG, individuals raised and sustained their contributions at high levels. Exclusion from the third WFG had no significant effect, but it reinforced increased contributions.
Willingness to incur personal cost to enhance the threat of exclusion	
H2a: Those who consider higher contributions as most socially appropriate will be more likely to assign exclusion tickets to those contributing less than that expectation.	**Supported**
H2b: Over the eight rounds of the pool exclusion PG game, there is negative or no relation between the number of exclusion tickets assigned and the number of exclusion tickets received by an individual.	**Not supported**
H2c: There is a negative relation between the total number of exclusion tickets assigned at the group level and the group average contribution levels.	**Supported**

5. Discussion

Financial mechanisms have been extensively studied in the economics literature as a way to overcome free-riding. In many experimental settings, however, social disapproval is bundled with financial incentives (i.e., financial punishment), making it difficult to disentangle whether behavioural responses reflect conformity to others' expectations, financial motives, or a mix of both. This confounding also makes it challenging to assess the long-term effectiveness of such mechanisms, which can diminish through counter-punishments, reduced overall group payoffs, and the loss of impact once financial incentives are removed.

Our study isolates the purely non-financial component of social disapproval by implement-

ing an endogenous social exclusion mechanism in a repeated public goods (PG) game. Our mechanism is designed to reflect real-world community interactions more closely by operationalising exclusion as a non-financial social consequence that arises endogenously from accumulated disapproval within a group. This stands in contrast to earlier studies on exclusion in PG games that linked it to financial incentives (Cinyabuguma et al., 2005; Davis & Johnson, 2015; Maier-Rigaud et al., 2010; Masclet, 2003). Our study is among the few (e.g., Charness & Yang, 2008) that examine an exclusion mechanism with the possibility for individuals to redeem themselves. This approach better reflects real social environments, where excluded individuals can modify their behaviour to regain social acceptance and rejoin group or organisational activities. Our exclusion mechanism reflects the time needed for such a social consequence to be imposed on a shared agreement, which is fundamental to the enforcement of normative expectations (Bicchieri, 2006).

Our design allows us to distinguish two channels: the effect of the threat of exclusion on PG game contributions, and the effect of actually experiencing exclusion on the threat's effectiveness. Exclusion in our experiment was from a word-formation game played within one's group. Our results show that, in the absence of financial incentives, a collective decline in contributions was successfully reversed by the mere threat of social exclusion. Individuals who received exclusion tickets responded by increasing their contributions in the following round. However, the marginal effect of this growing threat diminished over time, suggesting that the threat of social exclusion raised contributions early on, but the subsequent increases became less pronounced as the rounds progressed.

Actual exclusion was experienced mainly by individuals who consistently contributed below their group's average. Experiencing exclusion further heightened and reinforced the effectiveness of the threat. After the first exclusion, contributions rose on average by about 12.8 percent of the endowment and improved group coordination. The second exclusion reinforced this effect by maintaining the higher contribution levels. Only 5% of participants appeared relatively unresponsive to non-financial incentives, supporting the view that (the threat of) social exclusion was in this experiment an effective social mechanism for the vast majority.

Our results provide evidence that non-financial incentives alone can shape behaviour. In our experiment, participation in the group activity carried no direct or indirect financial consequences; yet exclusion still had a powerful effect. Participants' reactions to the threat of exclusion were driven entirely by the prospect of future non-financial social costs. In addition, because the exclusion mechanism was endogenously implemented by group members themselves, rather than imposed externally, it reflected the group's own normative expectations. This meant that participants were motivated to align with their peers' expectations in order to remain part of the group. This shows that the desire to take part in social (group)

activities with one's reference group can serve as a strong disincentive to free-ride, even when this activity lies outside the collective action problem itself and has no (in)direct financial consequences.

Previous research has already shown thatnon-financial social disapproval can foster cooperation (e.g., Brook & Servátka, 2016; Dugar, 2010; Gächter & Fehr, 1999; López-Pérez & Vorsatz, 2010; Masclet et al., 2003; Peeters & Vorsatz, 2013). More recently, Sparks et al. (2024) have argued that its long-term effectiveness depends on being linked to a credible social consequence. Our findings extend this literature by showing that social disapproval can increase its effectiveness among those initially less responsive when leveraged through a non-financial consequence, specifically, the threat and experience of exclusion.

Turning to how the exclusion mechanism itself emerges, our results contribute to our understanding of when and why social disapproval escalates within groups, and whether this escalation reflects communication and enforcement of normative expectations. As participants progressed through rounds of the PG game, they were increasingly willing to devote part of their payoffs to assign costly exclusion tickets. This dynamic was largely driven by individual normative expectations. Participants sanctioned others more when the others' contributions fell below their own perceived norm than when they matched it. They sanctioned above-norm contributions even less frequently. The fact that above-norm deviators received fewer exclusion tickets than non-deviators points to a shift toward more cooperative normative change, as shown in our parallel study (see Martinez-Felip et al., 2025).

Not only did the importance of perceived norms matter; their strength did too. Dannenberg et al. (2020) showed that participants' choice to implement a social exclusion mechanism is better explained by models incorporating social preferences than by the standard model of purely self-interested behaviour. Our results build on this and suggest that the endogenous emergence and stability of such a mechanism depend not only on the presence of cooperative normative expectations but also on their perceived strength. Participants with tight normative expectations were more likely than those with loose ones to enforce the norm when higher contribution levels became more dispersed. This pattern is consistent with prior work on norm strength linked with tolerance for deviant behaviour (Gelfand et al., 2017; Gelfand et al., 2011).

In a PG game without any possibilities for punishment, Dimant et al. (2025) show that contribution environments characterised by high variance, lead to greater dispersion in contributions. By contrast, we find that when groups are able or allowed to sanction inappropriate contributions, greater dispersion in contributions activates norm enforcement, particularly among those with tight normative expectations. This enforcement pushed behaviour toward higher contributions, reinforcing the emergence of socially beneficial expectations

(see Martinez-Felip et al., 2025). Although the primary driver was norm enforcement, about 13% of participants also appeared motivated by self-serving reasons: they assigned tickets to others to reduce their own chance of being excluded. This finding shows that the group activity employed in this study (the word formation game) fulfilled its purpose as an activity valued by the group, which in turn made the threat of exclusion a meaningful sanction.

These dynamics have a clear interpretation. As higher contributions become common and enforced, the expected social cost of free riding increases. Andreoni et al. (2021) and Janas et al. (2025) argue that normative change is often hindered by the costs of transitioning to a new norm, and that tipping points occur when perceived benefits exceed these costs and expectations of collective change reinforce the transition. Our results suggest that non-financial exclusion, when endogenous, can help overcome these frictions and move groups toward tipping points, facilitating norm emergence and long-lasting cooperation without relying on financial mechanisms.

Taken together, our results show that social exclusion, when endogenous, can trigger the emergence of coordinated, norm-consistent cooperation. This was evidenced by (1) more cooperative normative expectations (Martinez-Felip et al., 2025) that were enforced through assignment of exclusion tickets, and (2) both higher contributions and greater within-group coordination in contributions after exposure to the threat and experience of exclusion. These findings highlight the coupled dynamics between normative expectations, non-financial incentives that emerge within groups, and actual group behaviour in collective action.

A limitation of our work is that, although the exclusion mechanism did make norm-consistent behaviour emerge, the limited number of post-exclusion periods prevents us from drawing firm conclusions about longer-term effects. In particular, it is unclear whether the observed cooperation reflects genuine norm internalisation or only short-lived enforcement. We cannot tell whether contributions would persist without continued enforcement—that is, whether a tipping point was reached. Future research should extend group interactions and include repeated measures of personal norms alongside normative expectations, to capture the convergence between external (non-financial) enforcement and internalised adherence.

6. Conclusion

Extensive research on how communities address collective action problems shows that financial incentives can effectively promote cooperation. In particular, financial punishments for socially disapproved behaviour reduce the personal gains from free-riding and thereby sustain contributions (e.g., Balliet et al., 2011; Chaudhuri, 2011; Fehr & Gächter, 2000). However, we still lack a clear understanding of when and why social disapproval escalates within groups and what effect it has on contributions to the provision of public goods. Moreover, because

these mechanisms rely on financial incentives, it remains unclear whether behavioural responses to punishment are driven by compliance with normative expectations, by self-interest, or by both. As a result, the long-term effectiveness of policies that provide financial incentives is potentially compromised once those incentives are not in place (Brandts & Cooper, 2006; Hamman et al., 2007; Nakagawa et al., 2022).

It is well-established that communities often employ non-financial mechanisms to signal and enforce normative expectations (Bicchieri, 2006; Ostrom, 1990, 2000; Ullmann-Margalit, 1977). This study extends previous research on social disapproval and cooperative behaviour (e.g., Brook & Servátka, 2016; Dugar, 2010; Gächter & Fehr, 1999; López-Pérez & Vorsatz, 2010; Masclet et al., 2003; Peeters & Vorsatz, 2013) by considering its effectiveness when linked to a purely social consequence: the threat of exclusion from community interactions (Francis, 1985; Gruter & Masters, 1986; Kurzban & Leary, 2001; Ostrom, 1990; Ouwerkerk et al., 2005; Williams, 2001).

Our results make two contributions. First, non-financial incentives that emerged within groups and operated through a social exclusion mechanism provided a powerful disincentive to free-riding. The exclusion mechanism imposed a social cost that reversed the decline in contributions observed in the Baseline PG game, fostered coordination at higher contribution levels, and was effective even when only threatened. Social exclusion was mainly targeted at those contributing less than their group average, and experiencing it positively reinforced the effectiveness of the threat for those participants. Second, participants' willingness to incur personal costs to enhance the threat of exclusion increased over time and was shaped by their cooperative normative expectations. Crucially, this willingness increased when higher within-group contribution levels became more dispersed, but this effect was driven only by participants perceiving tight norms. Our findings therefore highlight the dynamic interplay between normative expectations, non-financial sanctions, and actual group behaviour. This interplay underscores a potential self-reinforcing cycle of collectively beneficial behaviours sustained by non-financial incentives.

Our findings point to the potential for a self-reinforcing feedback for cooperation. A promising avenue for future work is to explore how these dynamics are shaped by socio-cultural environments and social networks (Alesina & Giuliano, 2015; Gavrilets & Richerson, 2017; Gelfand et al., 2011; Hu et al., 2015). Finally, our results are important for advancing our theoretical understanding of the role of non-financial incentives in resolving collective action problems. They may also be relevant for policymakers seeking to promote long-lasting pro-social behaviours without relying on continuous financial support.

Data availability

The data and analysis files are publicly available on the project's OSF site: https://osf.io/v4q9b

Founding sources

This research was conducted as part of a PhD and received funding from the Australian Research Council, ARC-Discovery Project DP220100482. The experimental work was conducted using the University of Western Australia's facilities. We thank The University of Western Australia for providing the Scholarship for International Research Fees for the first author to undertake the PhD study.

References

Abadie, A., Diamond, A., & Hainmueller, J. (2010). Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program. *Journal of the American Statistical Association*, *105*(490), 493-505. https://doi.org/10.1198/jasa.2009.ap08746

Abadie, A., & Gardeazabal, J. (2003). The Economic Costs of Conflict: A Case Study of the Basque Country. *American Economic Review*, *93*(1), 113-132. https://doi.org/10.1257/000282803321455188

Alesina, A., & Giuliano, P. (2015). Culture and Institutions. *Journal of Economic Literature*, *53*(4), 898-944. https://doi.org/10.1257/jel.53.4.898

Andreoni, J., Nikiforakis, N., & Siegenthaler, S. (2021). Predicting social tipping and norm change in controlled experiments. *Proceedings of the National Academy of Sciences*, *118*(16), e2014893118-e2014893118. https://doi.org/10.1073/pnas.2014893118

Aycinena, D., Bogliacino, F., & Kimbrough, E. O. (2024). Measuring Norms: A Comparison of the Predictive and Descriptive Power of Three Methods. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4663919

Balliet, D., Mulder, L. B., & Van Lange, P. A. M. (2011). Reward, punishment, and cooperation: a meta-analysis. *Psychological Bulletin*, *137*(4), 594-594. https://doi.org/10.1037/a0023489

Barrett, S., & Dannenberg, A. (2016). An experimental investigation into 'pledge and review' in climate negotiations. *Climatic Change*, *138*(1), 339-351. https://doi.org/10.1007/s10584-016-1711-4

Bernstein, M. J., & Claypool, H. M. (2012). Not all social exclusions are created equal: Emo-

tional distress following social exclusion is moderated by exclusion paradigm. *Social Influence*, *7*(2), 113-130. https://doi.org/10.1080/15534510.2012.664326

Betsch, C., Böhm, R., Korn, L., & Holtmann, C. (2017). On the benefits of explaining herd immunity in vaccine advocacy. *Nature Human Behaviour*, *1*(3), 0056. https://doi.org/10.1038/s41562-017-0056

Bicchieri, C. (2006). *The grammar of society: The nature and dynamics of social norms*. Cambridge University Press. https://doi.org/10.1017/CBO9780511616037

Brandts, J., & Cooper, D. J. (2006). A Change Would Do You Good .... An Experimental Study on How to Overcome Coordination Failure in Organizations. *American Economic Review*, *96*(3), 669-693. https://doi.org/10.1257/aer.96.3.669

Brook, R., & Servátka, M. (2016). The anticipatory effect of nonverbal communication. *Economics Letters*, *144*, 45-48. https://doi.org/10.1016/j.econlet.2016.04.033

Charness, G., Cobo-Reyes, R., & Jiménez, N. (2014). Identities, selection, and contributions in a public-goods game. *Games and Economic Behavior*, *87*, 322-338. https://doi.org/10.1016/j.geb.2014.05.002

Charness, G., Gneezy, U., & Halladay, B. (2016). Experimental methods: Pay one or pay all. *Journal of Economic Behavior & Organization*, *131*, 141-150. https://doi.org/10.1016/j.jebo.2016.08.010

Charness, G. B., & Yang, C.-L. (2008). *Endogenous group formation and public goods provision: Exclusion, exit, mergers, and redemption*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=932251

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

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

Cinyabuguma, M., Page, T., & Putterman, L. (2005). Cooperation under the threat of expulsion in a public goods experiment. *Journal of Public Economics*, *89*(8), 1421-1435. https://doi.org/10.1016/j.jpubeco.2004.05.011

D'Adda, G., Drouvelis, M., & Nosenzo, D. (2016). Norm elicitation in within-subject designs: Testing for order effects. *Journal of Behavioral and Experimental Economics*, *62*, 1-7. https://doi.org/10.1016/j.socec.2016.02.003

Dannenberg, A., Haita-Falah, C., & Zitzelsberger, S. (2020). Voting on the threat of exclusion in a public goods experiment. *Experimental Economics*, *23*(1), 84-109. https://doi.org/10.1007/s10683-019-09609-y

Davis, B. J., & Johnson, D. B. (2015). Water cooler ostracism: Social exclusion as a punishment mechanism. *Eastern Economic Journal*, *41*(1), 126-151. https://doi.org/10.1057/eej.2014.2

Dimant, E. (2023). Beyond average: A method for measuring the tightness, looseness, and polarization of social norms. *Economics Letters*, *233*, 111417-111417. https://doi.org/10.1016/j.econlet.2023.111417

Dimant, E., Gelfand, M., Hochleitner, A., & Sonderegger, S. (2025). Strategic Behavior with Tight, Loose, and Polarized Norms. *Management Science*, *71*(3). https://doi.org/10.1287/mnsc.2023.01022

Dohmen, T., Huffman, D., Schupp, J., Falk, A., Sunde, U., & Wagner, G. G. (2011). Individual risk attitudes: Measurement, determinants, and behavioral consequences. *Journal of the European Economic Association*, *9*(3), 522-550. https://doi.org/10.1111/j.1542-4774.2011.01015.x

Dreber, A., Rand, D. G., Fudenberg, D., & Nowak, M. A. (2008). Winners don't punish. *Nature*, *452*(7185), 348-351. https://doi.org/10.1038/nature06723

Dugar, S. (2010). Nonmonetary sanctions and rewards in an experimental coordination game. *Journal of Economic Behavior & Organization*, *73*(3), 377-386. https://doi.org/10.1016/j.jebo.2009.11.003

Eriksson, K., Strimling, P., Andersson, P. A., & Lindholm, T. (2017). Costly punishment in the ultimatum game evokes moral concern, in particular when framed as payoff reduction. *Journal of Experimental Social Psychology*, *69*, 59-64. https://doi.org/10.1016/j.jesp.2016.09.004

Erkut, H. (2020). Incentivized Measurement of Social Norms Using Coordination Games. *Analyse Und Kritik*, *42*(1), 97-106. https://doi.org/10.1515/auk-2020-0004

Fallucchi, F., & Nosenzo, D. (2022). The coordinating power of social norms. *Experimental Economics*, *25*(1), 1-25. https://doi.org/10.1007/s10683-021-09717-8

Fehr, E., & Fischbacher, U. (2004). Social norms and human cooperation. *Trends in Cognitive Sciences*, *8*(4), 185-190. https://doi.org/10.1016/j.tics.2004.02.007

Fehr, E., & Gächter, S. (2000). Cooperation and punishment in public goods experiments. *American Economic Review*, *90*(4), 980-994. https://doi.org/10.1257/aer.90.4.980

Fehr, E., & Rockenbach, B. (2003). Detrimental effects of sanctions on human altruism. *Nature*, *422*, 137-140. https://doi.org/10.1038/nature01474

Fehr, E., & Schurtenberger, I. (2018). Normative foundations of human cooperation. *Nature Human Behaviour*, *2*(7), 458-468. https://doi.org/10.1038/s41562-018-0385-5

Fischbacher, U., & Gächter, S. (2010). Social Preferences, Beliefs, and the Dynamics of Free Riding in Public Goods Experiments. *American Economic Review*, *100*(1), 541-556. https://doi.org/10.1257/aer.100.1.541

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

Francis, H. (1985). The Law, Oral Tradition and the Mining Community. *Journal of Law and Society*, *12*(3), 267-267. https://doi.org/10.2307/1410120

Gächter, S., & Fehr, E. (1999). Collective action as a social exchange. *Journal of Economic Behavior & Organization*, *39*(4), 341-369. https://doi.org/10.1016/S0167-2681(99)00045-1

Gavrilets, S., & Richerson, P. J. (2017). Collective action and the evolution of social norm internalization. *Proceedings of the National Academy of Sciences*, *114*(23), 6068-6073. https://doi.org/10.1073/pnas.1703857114

Gelfand, M. J., Harrington, J. R., & Jackson, J. C. (2017). The Strength of Social Norms Across Human Groups. *Perspectives on Psychological Science*, *12*(5), 800-809. https://doi.org/10.1177/1745691617708631

Gelfand, M. J., Raver, J. L., Nishii, L., Leslie, L. M., Lun, J., Lim, B. C., Duan, L., Almaliach, A., Ang, S., & Arnadottir, J. (2011). Differences between tight and loose cultures: A 33-nation study. *Science*, *332*(6033), 1100-1104. https://doi.org/10.1126/science.1197754

Gruter, M., & Masters, R. D. (1986). Ostracism as a social and biological phenomenon: An introduction. *Ethology and Sociobiology*, *7*(3), 149-158. https://doi.org/10.1016/0162-3095(86)90043-9

Hamman, J., Rick, S., & Weber, R. A. (2007). Solving coordination failure with "all-or-none" group-level incentives. *Experimental Economics*, *10*(3), 285-303. https://doi.org/10.1007/s10683-007-9179-8

Hardin, G. (1968). The Tragedy of the Commons. *Science*, *162*(3859), 1243-1248. https://doi.org/10.1126/science.162.3859.1243

Herrmann, B., Thoni, C., & Gachter, S. (2008). Antisocial punishment across societies. *Science*, *319*(5868), 1362-1367. https://doi.org/10.1126/science.1153808

Hershey, J., Asch, D., Thumasathit, T., Meszaros, J., & Waters, V. (1994). The Roles of Altruism, Free Riding, and Bandwagoning in Vaccination Decisions. *Organizational Behavior and Human Decision Processes*, *59*, 177-187. https://doi.org/10.1006/obhd.1994.1055

Hox, J. J., Moerbeek, M., & van de Schoot, R. (2017). *Multilevel analysis: Techniques and applications*. Routledge.

Hu, H.-H., Lin, J., & Cui, W. (2015). Cultural differences and collective action: A social network perspective. *Complexity*, *20*(4), 68-77. https://doi.org/10.1002/cplx.21515

Janas, M., Nikiforakis, N., & Siegenthaler, S. (2025). Predicting norm change using threshold models. *Current Opinion in Psychology*, *62*, 101994-101994. https://doi.org/10.1016/j.copsyc.2025.101994

Kimbrough, E. O., & Vostroknutov, A. (2016). Norms make preferences social. *Journal of the European Economic Association*, *14*(3), 608-638. https://doi.org/10.1111/jeea.12152

Krupka, E. L., & Weber, R. A. (2013). Identifying social norms using coordination games: Why does dictator game sharing vary? *Journal of the European Economic Association*, *11*(3), 495-524. https://doi.org/10.1111/jeea.12006

Kurzban, R., & Leary, M. R. (2001). Evolutionary origins of stigmatization: the functions of social exclusion. *Psychological Bulletin*, *127*(2), 187-208. https://doi.org/10.1037/0033-2909.127.2.187

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

Li, K. T. (2019). Statistical Inference for Average Treatment Effects Estimated by Synthetic Control Methods. *Journal of the American Statistical Association*, *115*(532), 2068-2083. https://doi.org/10.1080/01621459.2019.1686986

López-Pérez, R., & Vorsatz, M. (2010). On approval and disapproval: Theory and experiments. *Journal of Economic Psychology*, *31*(4), 527-541. https://doi.org/10.1016/j.joep.2010.03.016

Lüdecke, D., Ben-Shachar, M. S., Patil, I., Waggoner, P., & Makowski, D. (2021). performance: An R Package for Assessment, Comparison and Testing of Statistical Models. *Journal of Open Source Software*, *6*(60), 3139-3139. https://doi.org/10.21105/joss.03139

Maier-Rigaud, F. P., Martinsson, P., & Staffiero, G. (2010). Ostracism and the provision of a public good: experimental evidence. *Journal of Economic Behavior and Organization*, *73*(3), 387-395. https://doi.org/10.1016/j.jebo.2009.11.001

Martinez-Felip, D., Schilizzi, S. G. M., & Nguyen, C. (2025). How does experienced behavior change normative expectations regarding socially beneficial actions? In: SocArXiv. https://doi.org/10.31235/osf.io/vy6z2_v5

Masclet, D. (2003). Ostracism in work teams: A public good experiment. *International Journal of Manpower*, *24*(7), 867-889. https://doi.org/10.1108/01437720310502177

Masclet, D., Noussair, C., Tucker, S., & Villeval, M.-C. (2003). Monetary and Nonmonetary Punishment in the Voluntary Contributions Mechanism. *The American Economic Review*, *93*(1), 366-380. https://doi.org/10.1257/000282803321455359

Nakagawa, M., Lefebvre, M., & Stenger, A. (2022). Long-lasting effects of incentives and social preference: A public goods experiment. *PLOS ONE*, *17*(8), e0273014. https://doi.org/10.1371/journal.pone.0273014

Nielsen, Y. A., & Pfattheicher, S. (2024). Separating individual and group-level cooperation in the Public Goods Game. *PNAS Nexus*, *3*(5), pgae200-pgae200. https://doi.org/10.1093/pnasnexus/pgae200

Ostrom, E. (1990). *Governing the commons: the evolution of institutions for collective action*. Cambridge University Press. https://doi.org/10.2307/3146384

Ostrom, E. (2000). Collective Action and the Evolution of Social Norms. *Journal of Economic Perspectives*, *14*(3), 137-158. https://doi.org/10.1257/jep.14.3.137

Ouwerkerk, J. W., Kerr, N. L., Gallucci, M., & Van Lange, P. A. M. (2005). Avoiding the Social Death Penalty: Ostracism and Cooperation in Social Dilemmas. In *The Social Outcast: Ostracism, Social Exclusion, Rejection, and Bullying* (pp. 321-332). Psychology Press.

Peeters, R., & Vorsatz, M. (2013). IMMATERIAL REWARDS AND SANCTIONS IN A VOLUNTARY CONTRIBUTION EXPERIMENT. *Economic Inquiry*, *51*(2), 1442-1456. https://doi.org/10.1111/j.1465-7295.2011.00433.x

Selten, R., & Stoecker, R. (1986). End behavior in sequences of finite Prisoner's Dilemma supergames A learning theory approach. *Journal of Economic Behavior and Organization*, *7*(1), 47-70. https://doi.org/10.1016/0167-2681(86)90021-1

Sparks, A., Burleigh, T., & Barclay, P. (2024). Expressed disapproval does not sustain long-term cooperation as effectively as costly punishment. *Evolutionary Human Sciences*, *6*. https://doi.org/10.1017/ehs.2024.41

Ullmann-Margalit, E. (1977). *The Emergence of Norms*. Oxford University Press.

Williams, K. D. (2001). *Ostracism: The Power of Silence*. Guilford Publications.

Williams, K. D. (2007). Ostracism. *Annual Review of Psychology*, *58*, 425-452. https://doi.

org/10.1146/annurev.psych.58.110405.085641

Xiao, E. (2013). Profit-seeking punishment corrupts norm obedience. *Games and Economic Behavior*, *77*(1), 321-344. https://doi.org/10.1016/j.geb.2012.10.010

Zadro, L., Williams, K. D., & Richardson, R. (2004). How low can you go? Ostracism by a computer is sufficient to lower self-reported levels of belonging, control, self-esteem, and meaningful existence. *Journal of Experimental Social Psychology*, *40*(4), 560-567. https://doi.org/10.1016/j.jesp.2003.11.006

Zelmer, J. (2003). Linear Public Goods Experiments: A Meta-Analysis. *Experimental Economics*, *6*(3), 299-310. https://doi.org/10.1023/A:1026277420119

Appendix

A. Threat of exclusion on reversing the decline in PG contributions

Table A.1 provides statistical support for differences in contribution levels between the Baseline and Pool exclusion PG games. Given the panel structure of the data, with repeated measures at the individual level, we employ mixed-effects regression models with random individual effects. To account for significant between-group differences and contribution tendencies, we cluster standard errors at the group level. *Table A.1: Mixed-effects estimates of PG contributions (out of 50 points)*

Dependent variable: PG contribution	**Coef.**	**Std. Err.**	**t-stat**	**p-value**
(Intercept)	15.224	9.079	1.677	0.120
Threat of exclusion (vs Baseline PG game)	0.449	2.010	0.223	0.825
Period	-1.542	0.518	-2.980	0.006^{***}
Threat of exclusion $\times$ Period	2.331	0.620	3.760	$< 0.001^{***}$

Periods considered include Rounds 1 to 10: the 5 rounds of the Baseline PG game and the first 5 rounds of the Pool Exclusion PG game, during which only the threat of exclusion was present. The *threat of exclusion* is represented by a dummy variable (1 = first 5 rounds of the Pool Exclusion PG game, 0 = Baseline PG game). Standard errors are clustered at the group level. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.
Additional controls: gender, demographics, risk, environmental attitudes, age.

A negative and significant coefficient of *period* indicates that contributions tended to decline over time during the five rounds of the Baseline PG game. In contrast, a positive and highly significant coefficient on *period* and *Threat of exclusion* interaction indicates that under the first five rounds of the Pool Exclusion PG game where only the threat of exclusion was present, this decline was not only offset but reversed, leading to increasing contributions over time.

B. Between-group and within-group variance evolution *Table B.1: Variance components and ICC for Contributions by Treatment and Period.*

Table 5: Intra-class correlations (ICC) over time by treatment

Treatment	Period	ICC	Between-group variance	Within-group variance
Baseline PG game	1	0.000	0.000	–
Baseline PG game	2	0.103	21.967	190.876
Baseline PG game	3	0.350	89.422	165.626
Baseline PG game	4	0.552	142.326	115.256
Baseline PG game	5	0.626	174.083	103.583
Pool Exclusion PG game	6	0.373	69.883	117.153
Pool Exclusion PG game	7	0.749	138.455	46.283
Pool Exclusion PG game	8	0.720	142.211	55.060
Pool Exclusion PG game	9	0.721	155.627	60.213
Pool Exclusion PG game	10	0.668	151.786	75.263
Pool Exclusion PG game	11	0.783	169.869	47.023
Pool Exclusion PG game	12	0.792	187.188	49.073
Pool Exclusion PG game	13	0.737	176.219	62.956

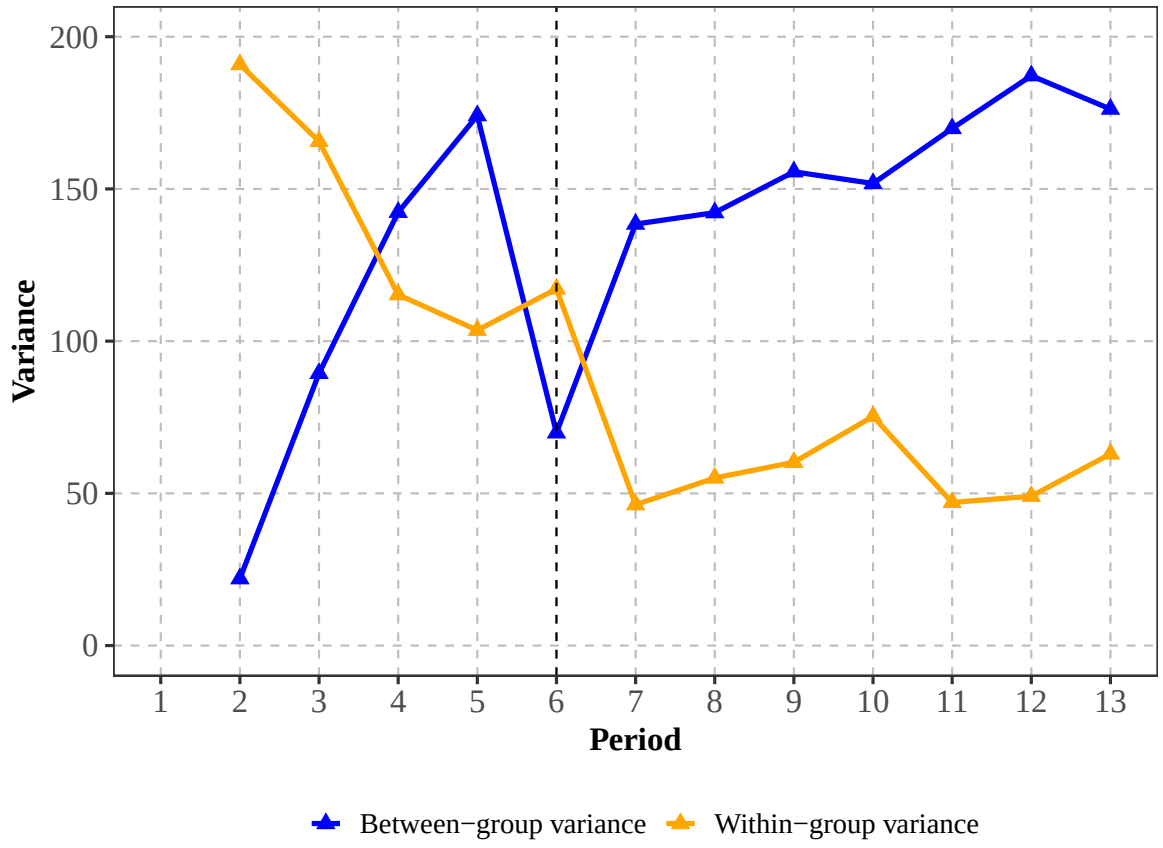

Figure B.1: Variance components evolution.

C. Synthetic control method – Parallel trends assumption for DiD estimator

The Synthetic Control Method (SCM) is a data-driven procedure that aims to estimate treatment effects in comparative case studies (Abadie et al., 2010; Abadie & Gardeazabal, 2003). Its goal is to construct a synthetic unit that serves as a reliable counterfactual for what would have happened if the treated unit had not been treated. This counterfactual is constructed by creating a convex combination of untreated units that closely replicates the treated unit's outcome evolution during the pre-treatment period. Treatment effects are then estimated by comparing the post-treatment outcomes of the synthetic control and the treated unit. Importantly, valid statistical inference of treatment effect estimates using the SCM relies on having a large number of pre-treatment and post-treatment observations (Li, 2019).

In our experiment, when examining the effect of experiencing exclusion from the second WFG, only five pre-exclusion periods (rounds 6 to 10) and two post-exclusion periods (rounds 11 and 12) were available to construct the synthetic excluded unit and estimate the effect of exclusion, respectively. This small number of post-exclusion observations limits the reliability of exclusion effect estimates derived from the SCM. Consequently, we use the SCM only as complementary evidence to evaluate the plausibility of the parallel trends assumption required for the Difference-in-Differences (*DiD*) estimator used to estimate the effect of exclusion (see subsection 4.2).

To construct the synthetic excluded unit for the second WFG, we created a synthetic counterpart for each excluded participant by using non-excluded participants as the donor pool. Figure C.1. displays the average actual contributions of participants excluded after the second WFG and the average contributions of their synthetic counterparts, both before and after exclusion.

We do not construct the synthetic excluded unit for the third WFG due to the insufficient number of pre-exclusion periods (rounds 11 and 12) and post-exclusion periods (only round 13).

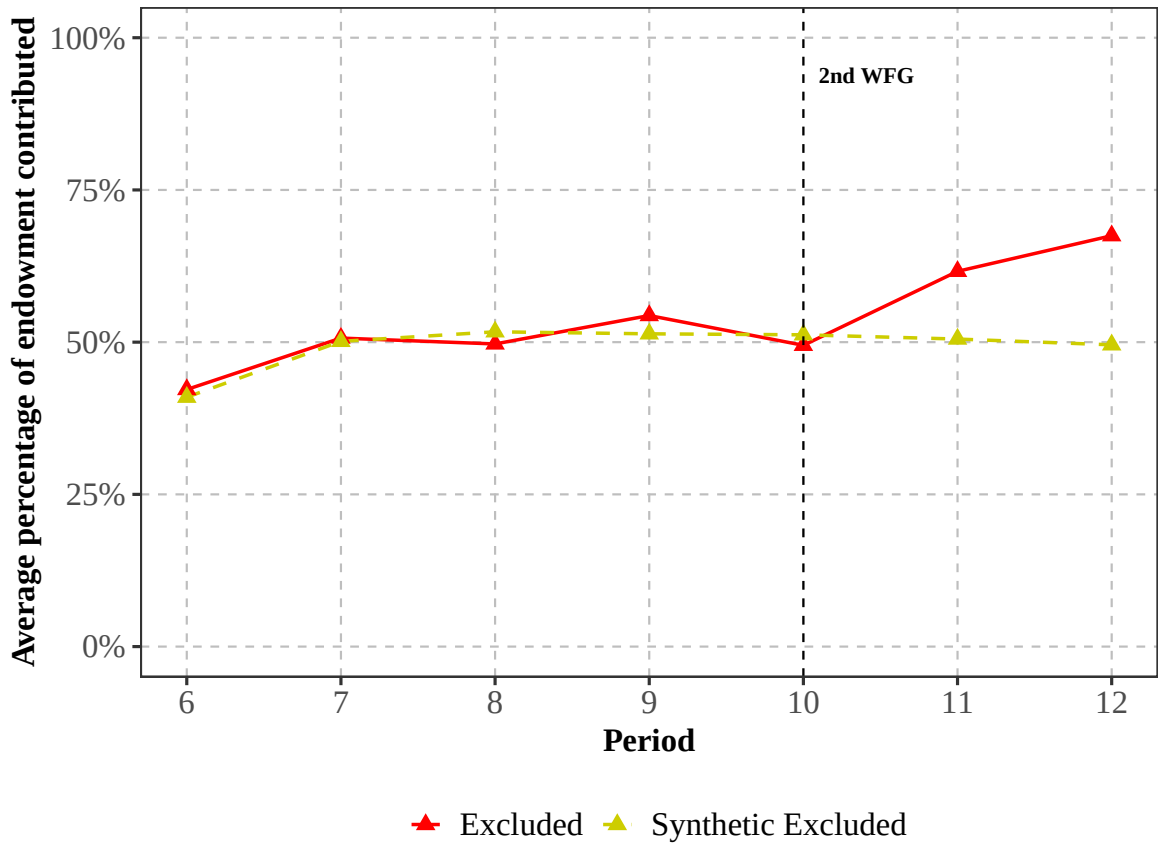

Figure C.1: Average PG contributions (in % of endowment) for Excluded participants and their Synthetic counterparts before and after the 2nd WFG.

D. Supplementary material. Experimental instructions