European Economic Review 164 (2024) 104682

Contents lists available at ScienceDirect

European Economic Review

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

# Cooperation, punishment, and group change in multilevel public goods experiments

Kasper Otten [a,c,*], Vincent Buskens [a], Wojtek Przepiorka [a], Boaz Cherki [b], Salomon Israel [b]

[a] Utrecht University, Department of Sociology, the Netherlands
[b] Hebrew University of Jerusalem, Department of Psychology, Israel
[c] Research and Data Centre (WODC), the Netherlands

ARTICLE INFO

*Keywords:*
Multilevel public goods
Cooperation
Punishment
Norms

ABSTRACT

Peer punishment is regarded as an important element in sustaining human cooperation for public good provision. Many behavioral experiments have shown that public good provision is higher if cooperation norms can be enforced by peer punishment. These experiments predominantly focus on single-group public goods, in which people have to choose between their private interests and the interests of their group. However, many societal problems comprise multilevel public goods problems, where multiple local groups are nested within a larger global group. We study experimentally how punishment affects cooperation and norms in multilevel public goods games. In our lab experiment, two local groups are nested within a larger global group. Participants have to choose between not contributing, contributing locally, and contributing globally. Local contributions would lead to a fragmented outcome where two separate local public goods are provided, whereas global contributions would lead to a unified global good that benefits all. Moreover, we study whether cooperation and punishment patterns depend on the type of public good participants are initially exposed to: single-group or multilevel. Participants either begin in a single-group public goods game and then shift to a multilevel public goods game or vice versa. We find that punishment is less effective in multilevel public goods games than in single-group public goods games. Punishment only promotes cooperation in multilevel public goods games if people have prior experience with solving single-group public goods games. Our results refine the boundary conditions for the effectiveness of punishment and suggest that 'starting small' by first solving single-group public goods problems is helpful for successful multilevel public good provision.

## 1. Introduction

Human cooperation to provide public goods is key for the success of social groups ranging from families and work teams to nations and international organizations. Cooperation for public good provision often presents a social dilemma – contributing is individually costly but brings benefits for the group. This means that there are incentives to free-ride on others' contributions, but the public good is not provided if everybody free-rides (Olson, 1965). Our ancestors had to overcome such free-rider incentives when they hunted large

* Corresponding author at: Padualaan 14, 3584 CH Utrecht, the Netherlands.
*E-mail address:* k.d.otten@uu.nl (K. Otten).

https://doi.org/10.1016/j.euroecorev.2024.104682
Received 23 January 2023; Received in revised form 16 January 2024; Accepted 19 January 2024
Available online 20 January 2024

animals, shared food, or engaged in warfare (Bowles and Gintis, 2013; Hawkes et al., 1993; Hill, 2002). Nowadays, people face free-rider incentives in solving team production tasks, when paying taxes, or when taking measures to reduce their carbon footprint (Kallhoff, 2014). How people cooperate despite the incentives to free-ride is still a major scientific question. Peer punishment is often suggested as a solution to achieving cooperation for public good provision (Balafoutas and Nikiforakis, 2012; Balafoutas et al., 2014; Fehr and Gächter, 2002; Fehr and Schurtenberger, 2018; Henrich et al., 2006). If cooperation norms can be enforced through punishment, free-riding is discouraged and cooperation becomes more likely. A large number of behavioral experiments show that giving people the opportunity for peer punishment indeed promotes cooperation (Balliet et al., 2011; Chaudhuri, 2011) and group welfare in the long run (Gächter et al., 2008).

Previous experiments on the effect of punishment on cooperation typically focus on single-group public goods problems, where people have to choose between not cooperating or cooperating with their own group. However, in many real-life instances of public good provision, there are multiple local groups nested within a larger global group. Such public goods have been labeled multilevel public goods (sometimes also referred to as nested social dilemmas) (Aaldering and Böhm, 2020; Aaldering et al., 2018; Blackwell and McKee, 2003; Böhm et al., 2014; Buchan et al., 2011, 2009; Espinosa et al., 2019; Fellner and Lünser, 2014; Gallier et al., 2019; Israel et al., 2012; Lange et al., 2022; Polzer et al., 2009, 1999; Wit and Kerr, 2002). In multilevel public goods problems, individuals have to choose to what extent they act in their private interests, cooperate with their own local group (local cooperation), and cooperate with the larger global group (global cooperation). For example, employees are clustered in teams that are themselves clustered in departments or organizations and have to choose their effort for each level. Because multiple group memberships are increasingly the rule rather than the exception, interest in multilevel public goods has been growing (Aaldering et al., 2018; Aaldering and Böhm, 2020; Buchan et al., 2009, 2011; Espinosa et al., 2019). However, we do not yet know whether and how peer punishment promotes cooperation in multilevel public goods problems. The first aim of this study is to experimentally test the effect of peer punishment on cooperation in multilevel public goods problems.

Examples of multilevel public goods problems can be found in several domains. In charitable giving, people need to decide how much to donate and whether to donate to local charities or global charities. For national public goods such as social benefits, native populations need to decide whether and how they grant access to other groups such as immigrants (Degen et al., 2019). Scholars have also suggested that pressing international issues such as tackling the climate crisis and the COVID-19 pandemic involve elements of multilevel public goods problems (Buchan et al., 2011, 2009; Romano et al., 2021; Tavoni et al., 2011). In these examples, global cooperation can be impeded by individuals' tendency to pursue local interests or private interests. In companies, merging failures have been attributed to difficulties that employees face in collaborating across departments or divisions (Weber and Camerer, 2003). In charitable giving, people generally donate more to domestic causes than global causes, even when donations to global causes can do a lot more good (Grimson et al., 2020). In two-party political systems, parties sometimes pursue forms of public good provision that benefit their own supporters instead of the general public, leading to polarization and suboptimal policies (Dimant, 2022; Schultz, 1996). During the COVID-19 pandemic, several countries bought up a disproportionally large share of vaccines for their own populations, even though a more equitable global distribution would have been more effective to limit virus mutations that evade the vaccines (Ye et al., 2022). Finally, in combatting climate change, countries regularly take actions that reduce the negative environmental consequences for their local population but not for the global population (e.g., toxic waste trading) (Cotta, 2020).

Prior experiments on multilevel public goods problems show that also in the lab, global cooperation is often impeded by a tendency for local cooperation, even when global cooperation is collectively more beneficial (Aaldering and Böhm, 2020; Aaldering et al., 2018; Blackwell and McKee, 2003; Böhm et al., 2014; Buchan et al., 2011, 2009; Espinosa et al., 2019; Fellner and Lünser, 2014; Gallier et al., 2019; Israel et al., 2012; Lange et al., 2022; Polzer et al., 2009, 1999; Wit and Kerr, 2002). These experiments employ the multilevel public goods game (sometimes also referred to as a nested social dilemma), in which usually two local groups are nested within a larger global group. Participants then have to decide whether they want to not contribute at all, contribute to their local group, or contribute to a global good that benefits all. The multilevel public goods game is considered a central paradigm to study social fragmentation (Van Dijk and De Dreu, 2021). If people contribute locally instead of globally, a fragmented outcome ensues in which two separate local public goods are provided instead of a unified global good. The extent to which people prefer to cooperate locally or globally depends on the returns to both types of cooperation (Blackwell and McKee, 2003; Böhm et al., 2014; Fellner and Lünser, 2014; Gallier et al., 2019), the observability of both types of cooperation (Fellner and Lünser, 2014), resource inequality (Lange et al., 2022), framing (Polzer et al., 1999), and social categorization (Wit and Kerr, 2002). Similar to single-group public goods experiments, cooperation decays over time in multilevel public goods experiments, and the decay is especially noticeable for global cooperation (Blackwell and McKee, 2003; Fellner and Lünser, 2014). The well-known solution to prevent cooperation decay in single-group public goods experiments is peer punishment (Balliet et al., 2011; Chaudhuri, 2011), but this solution has not yet been tested for multilevel public goods experiments.

Theoretically, punishment may be less effective in promoting cooperation in multilevel public goods problems than in single-group public goods problems. The reason is that the potential for normative disagreement is larger in multilevel public goods problems. In single-group public goods problems, the only way to cooperate is to contribute to the group and most agree that this is the appropriate thing to do (Cubitt et al., 2011; Kimbrough and Vostroknutov, 2016; Reuben and Riedl, 2013). Hence, the cooperation norm is clear, which facilitates the use of peer punishment to effectively enforce norms. Indeed, previous experiments suggest that punishment is only socially beneficial if complemented by strong cooperation norms (Bicchieri et al., 2021; Herrmann et al., 2008). In multilevel public goods problems, there are several plausible cooperation norms; local cooperation, global cooperation, or a mix between the two (Catola et al., 2021). Consequently, there is more ambiguity about the cooperation norm, creating the potential for disagreement (Nikiforakis et al., 2012; Otten et al., 2020; Rauhut and Winter, 2017; Winter et al., 2012, 2018; Wit and Kerr, 2002). Individuals who do not cooperate and are subsequently punished may not always know whether the punishment is meant to induce them to cooperate

locally, globally, or a mix between the two. As a result, they may not react with the type of cooperation that the punisher intended to instigate. Further punishment or a decay in cooperation in general may be the result. Individuals who cooperate at a certain level (e.g., locally or globally) because they believe this to be appropriate may be reluctant to change their behavior when being punished (Rauhut and Winter, 2017). In sum, punishment may be less effective in enforcing cooperation norms when the norm that is to be enforced is more ambiguous or disputed.

Even if punishment promotes cooperation in multilevel public goods problems, there is no guarantee that it leads to a collectively efficient form of cooperation. In multilevel public goods problems as strictly defined, global cooperation is more efficient than local cooperation, which in turn is more efficient than no cooperation at all (Blackwell and McKee, 2003; Böhm et al., 2014; Buchan et al., 2011, 2009; Gallier et al., 2019; Lange et al., 2022; Wit and Kerr, 2002). Hence, if punishment promotes local cooperation over global cooperation, it leads to a collectively inefficient and fragmented form of cooperation. As mentioned, previous experiments suggest that people indeed tend to cooperate more with ingroup members than outgroup members (Aaldering and Böhm, 2020; Aaldering et al., 2018; Balliet et al., 2014; Blackwell and McKee, 2003; Böhm et al., 2014; Buchan et al., 2009; Fellner and Lünser, 2014; Gallier et al., 2019; Israel et al., 2012; Lange et al., 2022; Polzer et al., 2009, 1999; Wit and Kerr, 2002). Moreover, evidence from intergroup experiments suggests that punishment benefits ingroup members more than outgroup members (Bernhard et al., 2006). Based on prior research showing that punishment promotes cooperation (Balliet et al., 2011; Chaudhuri, 2011), we preregistered the hypothesis that total cooperation (combining local and global cooperation) in multilevel public goods problems is higher when punishment is possible (H1). Exploratively, we examine what type of cooperation, if any, is increased by punishment (local versus global). We also assess whether the effect of punishment in multilevel public goods problems is different from that in single-group public goods problems.

Previous multilevel public goods experiments impose the multilevel structure from the start (Aaldering and Böhm, 2020; Aaldering et al., 2018; Blackwell and McKee, 2003; Böhm et al., 2014; Buchan et al., 2011, 2009; Espinosa et al., 2019; Fellner and Lünser, 2014; Gallier et al., 2019; Israel et al., 2012; Lange et al., 2022; Polzer et al., 2009, 1999; Wit and Kerr, 2002). However, in society, multilevel structures often arise over time from membership changes between local groups. That is, people start in local groups with single-group public goods problems but over time end up in a multilevel public goods problem when groups become more mixed due to migration. For example, many nations have become more ethnically diverse through immigration over the past several decades, increasing the need for collective multicultural cooperation. Companies may also become more multilevel over time through changes in their hierarchical structure and mergers. Whether the multilevel public goods problem is present from the start or arises after an initial single-group public goods problem may be crucial for the cooperation norms that emerge and hence the norms that are enforced through peer punishment. People who were initially in a single-group public goods problem may have developed norms of local public goods provision (Bernhard et al., 2006; Choi et al., 2019, 2021; Otten et al., 2022; Titlestad et al., 2019). The presence of these norms of local public good provision may hamper global public good provision when the multilevel public goods problem arises. That is, individuals may stick to their norms of local public good provision even when they enter a multilevel public goods problem in which global cooperation is collectively more beneficial. Previous experiments indeed suggest that norms are sticky and spill over to new settings (Andreoni et al., 2021; Duffy and Lafky, 2021; Efferson and Vogt, 2018; Engl et al., 2021; Guala and Mittone, 2010; Otten et al., 2021; Peysakhovich and Rand, 2013; Przepiorka et al., 2022; Smerdon et al., 2019). In contrast, people who are in a multilevel public goods problem without prior experience in a single-group public goods problem may not be constrained by prior norms of local public good provision and therefore more easily coordinate on global cooperation. We therefore hypothesized that global cooperation will be higher when the multilevel public goods problem is present from the start than when it arises after an initial single-group public goods problem (H2). Both H1 and H2 were preregistered.

Contrary to expectations, we find that punishment does not promote cooperation in multilevel public goods problems if groups do not have prior experience in single-group public goods problems. In these groups, there is a mix of local and global cooperation and a decay in global cooperation even with punishment. However, in groups that do have prior experience with single-group public goods problems, punishment does promote and sustain cooperation in multilevel public goods problems. What is more, punishment leads these groups to develop and enforce collectively efficient norms of global cooperation.

Multilevel public good problems present a common but understudied setting in which groups consist of members with mixed affiliations instead of a single affiliation. Our results show that efficient cooperation can be difficult to achieve when group members have mixed affiliations. Indeed, an often-studied mechanism – peer punishment – fails to promote efficient outcomes in some of our conditions involving such multilevel public goods with multiple possible cooperation norms. However, our results also convey an optimistic message, as experience with cooperation norms in groups composed of members with a single affiliation carries over to groups composed of members with mixed affiliations. Efficiency norms developed among group members with a single affiliation help mixed groups to coordinate on efficient cooperation norms and the effective use of peer punishment.

The paper proceeds as follows. Section 2 describes the methods, including the experimental design, measurements, procedures, and analyses. Section 3 presents the empirical results, and Section 4 contains a discussion of the results.

## 2. Methods

### 2.1. Design

We conduct a behavioral laboratory experiment with 220 participants. Using a repeated public goods game (PGG) (Fehr and Gächter, 2000), we vary whether participants can use peer punishment to enforce norms and whether they move from a single-group PGG to a multilevel PGG or vice versa. The experimental design is presented in Fig. 1. At the start of the experiment, we randomly assign participants their local group memberships via three colors: blue, red, or green[1]. The color of each participant is fixed throughout the experiment. Participants are then placed into groups of four members facing either a single-group or a multilevel PGG. In single-group PGGs, all four participants belong to the same local group (i.e., the same color). In multilevel PGGs, the four participants are divided into two local groups with two members each (e.g., two blue participants and two red participants). In each round of the game, participants receive an endowment of 20 monetary units (MU) and have to allocate this endowment between a private account that benefits only themselves, a local account that benefits only members of their own local group (local cooperation), and a global account that benefits all four members regardless of which local group they belong to (global cooperation).

Individual marginal returns are highest for the private account (1), followed by the local account (0.7), and the global account (0.5). We call groups consisting of members who can contribute to the same local account local groups, and we call groups consisting of members who can contribute to the same global account global groups. When a statement refers to all members regardless of which of the two accounts they share, we may use the term 'group' in a general sense.

In the single-group PGG (where all four members belong to the same local group), it is best for the individual to not contribute at all whereas it is best for the collective (all four members combined) to contribute to the local account. This can be seen from a numerical example. For every participant who keeps the 20 MU, they get 20 MU themselves, while all other members of the group receive nothing (collective payoff = 20 MU). For every participant who contributes the 20 MU to the local good, they get $20 \times 0.7 = 14$ MU themselves, and the three other members of the local group also each receive 14 MU (collective payoff = 56 MU). For every participant who contributes the 20 MU to the global account, they get $20 \times 0.5 = 10$ MU themselves, and the three other members of the local group also each receive 10 MU (collective payoff = 40 MU). Hence, not contributing is best for each individual participant (20 MU vs 14 MU when contributing locally vs 10 MU when contributing globally), but contributing locally is the best for the collective (56 MU vs 40 MU when contributing globally vs 20 MU when not contributing at all).

In the single-group PGG, the dilemma is thus between not contributing and contributing to the local account. Here, local contributions should thus not be seen as a preference for the local group over the global group (as these groups are congruent), but rather as the efficient form of cooperation. Despite global contributions being neither individually nor collectively efficient in single-group PGGs, we still provide participants the option to contribute globally. In this way, we do not create artificial differences between the single-group and multilevel PGGs in terms of the complexity of the decision situation. However, as we will see, virtually nobody contributes globally in single-group PGGs. The tension between not contributing, contributing locally, and contributing globally only arises when multiple local groups are nested within a global group. This is what we capture in the multilevel PGG, where the global group of four members consists of two local groups with two members each.

In the multilevel PGG, it is best for the individual to not contribute at all whereas it is best for the collective (all four members combined) to contribute to the global account. This can again be seen from a numerical example. For every participant who keeps the 20 MU, they get 20 MU themselves, while all other members of the group receive nothing (collective payoff = 20 MU). For every participant who contributes the 20 MU to the local account, they get 20 MU $\times$ 0.7 = 14 MU themselves, the other member of the same local group also receives 14 MU, but the two members from the other local group receive nothing (collective payoff = 28 MU). For every participant who contributes the 20 MU to the global account, they get 20 MU $\times$ 0.5 = 10 MU themselves, and all three other members of the group (both from the same local group and from the other local group) also each receive 10 MU (collective payoff = 40 MU). So the individual payoff is highest when not contributing, followed by contributing locally, followed by contributing globally (20 MU vs 14 MU vs 10 MU). The collective payoffs follow the exact opposite order; they are highest when contributing globally, followed by contributing locally, followed by not contributing (40 MU vs 28 MU vs 20 MU). Finally, the payoffs of one's local group member are highest when contributing locally followed by contributing globally and not contributing (14 MU vs 10 MU vs 0 MU).

Participants first play 10 rounds of either a single-group PGG (all four members of the same local group, Fig. 1A and B) or a multilevel PGG (two local groups with two members each, Fig. 1C and D) within their group. After the first 10 rounds, we replace two members per group with two members of another group in such a way that single-group PGGs become multilevel PGGs (e.g., in a group with four blue members, two of the blue members are replaced by two red members) and vice versa (e.g., in a group with two blue and two red members, the two red members are replaced by two blue members). This means that participants remain with one of their members from before the group change. We inform participants of this group change, and then let the reshaped groups play another set of 10 rounds. To reduce endgame effects, we tell participants that the precise number of rounds of each part remains unknown to them and falls between 8 and 12. Participants can see the color of the co-participants they interact with every round (i.e., to which local group they belong).

Because we vary whether participants start with a single-group or multilevel PGG and whether punishment is possible, we have 2 $\times$

[1] When the number of groups in a session is uneven, we need three colors in order to switch members between groups such that each reshaped group consists of two members of one color and two members of another color. Half of the sessions had an uneven number of groups, for the other sessions we only used red and blue colors.

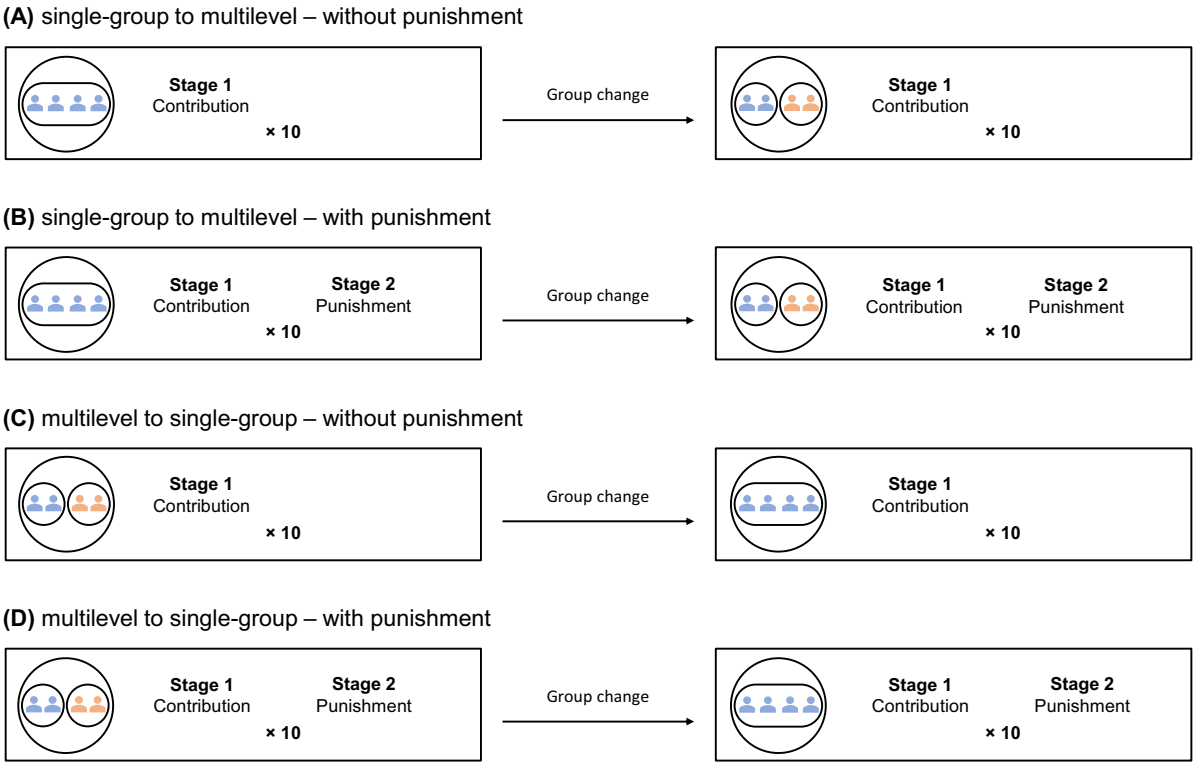



**Fig. 1.** Experimental design. Participants either start with a single-group public goods game (PGG) and after group change move to a multilevel PGG (*single-group to multilevel*) or vice versa (*multilevel to single-group*). We additionally manipulate whether peer punishment is possible. In both the single-group and multilevel PGG, participants play 10 rounds among the same four members. In each round, participants receive an endowment of 20 monetary units and have to allocate this endowment between a private account that benefits only themselves, a local account that benefits only members of their local group, and a global account that benefits all members regardless of which local group they belong to. Individual marginal returns are highest for the private account (1), followed by the local account (0.7), and then the global account (0.5). The accounts and conversion rates are available at all times.

$2 = 4$ conditions (Fig. 1, 52 participants in *A*, 52 participants in *B*, 56 participants in *C*, and 60 participants in *D*). We refer to the conditions in which people move from single-group PGGs to multilevel level PGGs as *single-group to multilevel* (Fig. 1A and B), and the conditions with the opposite order as *multilevel to single-group* (Fig. 1C and D). After each round, participants can see how much each of their group members contributed to the local and global accounts. In the punishment conditions (Fig. 1B and D), participants can then assign each other punishment points. Each assigned punishment point costs the punisher 1 MU and the punished member 3 MU. For example, if a participant assigns another member two punishment points, the participant loses 2 MU herself and the punished group member loses $2 \times 3$ MU $= 6$ MU. As is common in related studies (Fehr and Gächter, 2000; Reuben and Riedl, 2013), participants see how many punishment points they received, but not by whom. This curbs punishment driven by revenge motives instead of by contribution behavior and thereby helps to interpret punishment as a way to enforce norms. Participants can assign up to 10 punishment points to each group member per round, regardless of their earnings from the contribution stage. The punishment points were subtracted from participants' overall earnings at the end of the round. This could lead to negative earnings in a round, which occurred in 6 out of the 4400 cases (220 participants $\times$ 10 rounds in each part $\times$ 2 parts). Over the entire experiment, however, we guaranteed a minimum payment of 5 euros (as it turns out, the participant with the lowest payout earned 8.50 euros).

### 2.2. Norm measurements

Before round 1, we measured participants' personal normative views. To do so, we showed participants two hypothetical groups of four members, one in a single-group PGG (all four members the same color) and one in a multilevel PGG (two members with one color and two members with another color). We ask participants to report on their personal normative views for each of the two hypothetical groups with the question "In your view, what is the appropriate amount that each member should contribute to the color [local] account and the collective [global] account?". Participants could indicate a contribution to the color (local) and collective (global) account for each of the four members between 0 and 20. Participants can try out different combinations of contributions, and see how they affect the earnings of each group member in the two hypothetical groups (screenshots are provided in the Supplementary Material Figures S5 and S6).

Personal normative views are measured again before the 10th round of the first part of the experiment and also before the 10th round of the second part. Directly after these two measurements, we also asked participants to guess the personal normative views submitted by their group members. This measures participants' normative expectations rather than their own personal normative views (screenshots are provided in the Supplementary Material Figures S7 and S8). The measure of normative expectations is incentivized; if participants correctly guess the most frequent personal normative views among their group members, they receive a payment of 100 MU (~ 1.40 euros). The payment is separate for participants' guesses about their group members' views concerning (1) single-group PGGs and (2) multilevel PGGs. Hence, they can earn up to 200 MU (~ 2.80 euros) per measurement of normative expectations (400 MU in total because there are two measurement moments of normative expectations). The measurement of normative expectations is similar to prior studies (Krupka and Weber, 2013; Otten et al., 2021; Przepiorka et al., 2022; Reuben et al., 2015). Measurements of normative expectations were followed by measurements of empirical expectations, where participants guess what each of their group members will contribute to the local and global account in the upcoming round (a screenshot is provided in Figure S9). These expectations are also incentivized. Per participant, we randomly pick a guess about one of their three group members' contributions and compare this with the actual group member's contribution. If the guess and contribution are the same, participants receive a payment of 100 MU.

Because normative expectations are highly correlated with both personal normative views (0.83, $p < .001$) and empirical expectations (0.85, $p < .001$), we do not analyze them separately. We focus on normative expectations, as these are central in most accounts of norms (Bicchieri and Chavez, 2010; Kimbrough and Vostroknutov, 2016; Krupka and Weber, 2013; Otten et al., 2021; Przepiorka et al., 2022; Reuben et al., 2015). We did not elicit normative expectations already before round 1 because previous research suggests that group norms gradually develop through a dynamic learning process as group members interact with each other over time (Ostrom, 2000; Young, 2015).

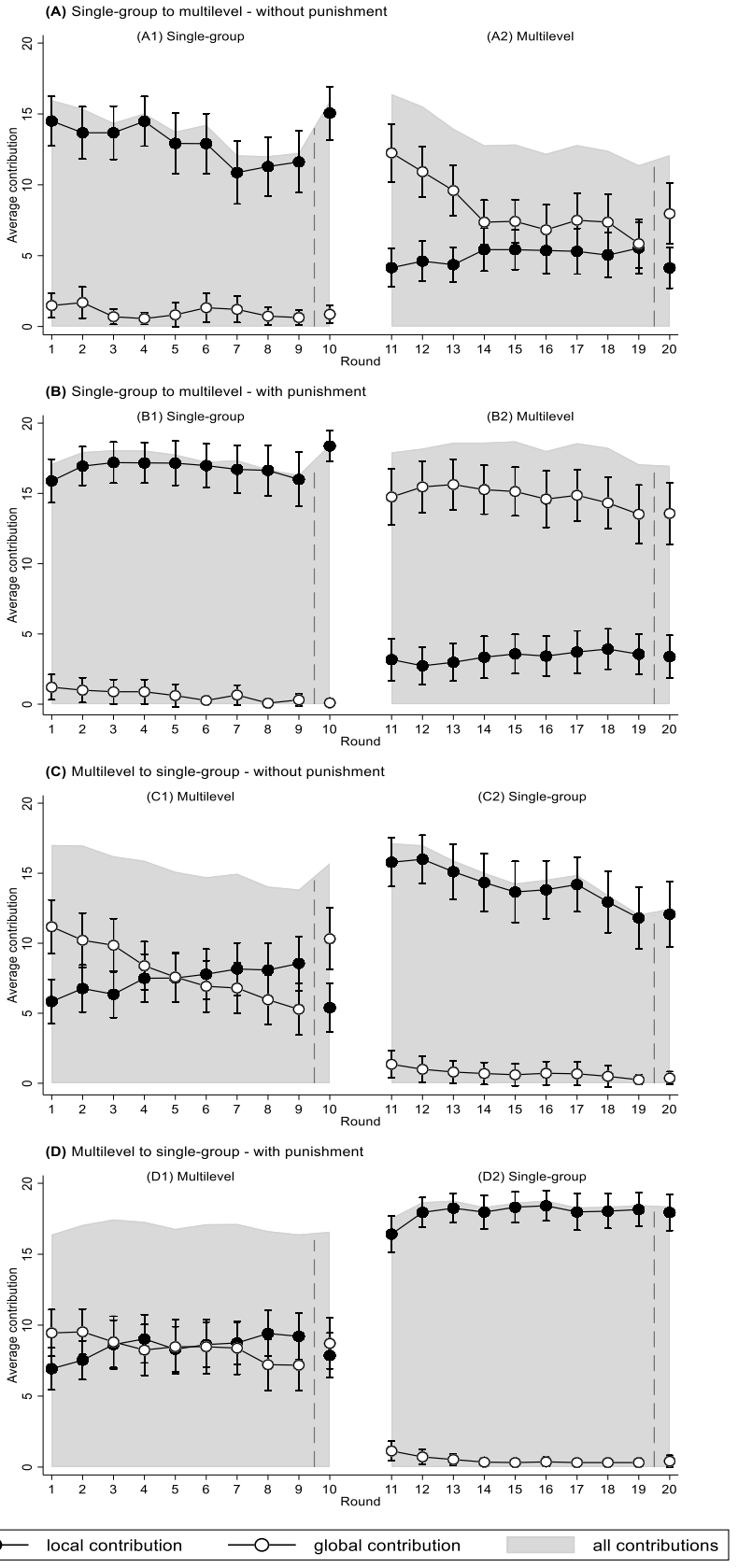
Because norms are measured directly before round 10, they can have an influence on contribution behavior in round 10. Indeed, prior research suggests that measuring norms can make them temporarily more salient and thereby increase cooperation (D'Adda et al., 2016; Krupka and Weber, 2009). As we will see, we indeed see an increase in cooperative behavior in round 10 compared to round 9 in some conditions, especially after the first set of norm measurements (before round 10 of the first set of rounds). In single-group PGGs, it is most cooperative to contribute locally (in terms of maximizing collective payoffs), and we see an increase in local contributions after the norm measurements in some conditions. In multilevel PGGs, it is most cooperative to contribute globally (in terms of maximizing collective payoffs), and we see an increase in global contributions after the norm measurements in some conditions. Although norm elicitations may affect behavior, they are important to us because the mechanisms underlying our hypotheses explicitly involve norms. By measuring normative expectations at the end of each part (single-group PGG and multilevel PGG) we can measure which norms developed and assess to what extent our supposed mechanisms are indeed at play. It would be problematic, for example, to only elicit norms at the end of the entire experiment, as people's norms may change throughout the experiment as groups are reshaped and they gain experience with new games (single-group vs multilevel PGG).

### 2.3. Procedures

We conducted the experiment at the Experimental Laboratory for Sociology and Economics (ELSE) at Utrecht University in October 2021. The experiment was programmed with z-Tree software (Fischbacher, 2007) and we recruited participants using the internet recruitment system ORSEE (Greiner, 2015). We ran 14 sessions with about 16 participants per session on average, obtaining a total of 220 participants. Sessions lasted about 1 h and 15 min. Payment depended on choices during the experiment and chance. Participants earned about 17 euros on average (min = 8.5, max = 22.5). Almost all participants were taking courses at Utrecht University. Out of the 220 participants, 114 were Dutch and 106 were from various other countries. Participants were on average 24 years old, 153 participants were female, 63 were male, and 4 identified as another gender. The sample size was selected to be able to detect medium effect sizes (Cohen's $d = 0.5$). We regard this effect size as a reasonable benchmark, given that meta-analyses on related studies using PGGs have found this effect size on average (Balliet et al., 2011; Spadaro et al., 2022).

Upon arrival in the lab, participants were randomly allocated to individual cubicles and told that they should not communicate with other participants. They were informed about the experiment through written instructions, which are provided in the Supplementary Material section S7. The written instructions inform the participants that the experiment is divided into two parts. The first part is explained in the written instructions, instructions on the second part are provided on the computer screen after the first part ends. After reading the instructions on the first part, participants had to complete a quiz to test their understanding of the instructions. 181 out of 220 participants answered all questions correctly on the first try, 27 participants answered 80–83 % correctly on the first try, 8 participants answered 60–67 % correctly on the first try, and 4 participants answered <60 % correctly on the first try. Participants were shown which questions they had answered incorrectly and had to correct these before they could continue with the experiment. The quiz is provided in the Supplementary Material section S7.

At the end of the experiment, participants also rated their own understanding of the experimental instructions. 192 participants reported a good understanding, 26 reported not bad, not good, and 2 reported bad. We did not exclude any participants or data points from the analyses to preserve random assignment. After the experiment, we presented participants with some hypothetical contribution scenarios and asked them to rate how appropriate these contributions were. We also asked them for some sociodemographic characteristics. These post-experiment measures are not analyzed in the current study. The experiment was preregistered before data collection at Open Science Framework: https://osf.io/5kawt. Informed consent was obtained from all participants and the experimental procedures were approved by the Faculty Ethical Review Board of the Faculty of Social and Behavioral Sciences of Utrecht University. All research was in line with relevant regulations. All data are openly available at the Open Science Framework: https://




*(caption on next page)*

**Fig. 2.** Cooperation in single-group and multilevel public goods games. We present the 10 contribution decisions in the single-group and multilevel public goods game for each condition. Mean contributions are indicated via markers and 95 % confidence intervals via capped spikes. The vertical dashed lines between rounds 9 and 10 and between rounds 19 and 20 indicate that norm measurements occurred between these rounds. In multilevel public goods games, we find that punishment increases global cooperation when participants encountered a single-group public goods game before the multilevel public goods game (panel B2 vs A2). Punishment does not increase global cooperation when participants encountered the multilevel public goods game before the single-group public goods game (panel D1 vs C1). In single-group public goods games, we find that virtually all cooperation is local cooperation and that local cooperation is higher if peer punishment is possible (panels B1 vs A1, and D2 vs C2).

doi.org/10.17605/OSF.IO/2DKGP.

### 2.4. Analyses

We run linear regression models to estimate whether behavior and normative perceptions depend on experimental conditions and the PGG (single-group or multilevel). For analyses that have repeated observations within individuals and groups, we run three types of regression models. The first type accounts for repeated observations by estimating individual-cluster robust standard errors. The second type accounts for repeated observations by estimating individual-level and group-level random effects in multilevel (panel) regressions. The third type accounts for repeated observations by estimating group-cluster robust standard errors. Because the results of the three types of models are substantively similar, we only report on regression models with individual-cluster robust standard errors in the main text. The multilevel (panel) regression models can be found in Supplementary Material section S4 (Tables S13 through S17), the linear regression models with group-cluster robust standard errors in Supplementary Material section S5 (Tables S18 through S22). Finally, we repeat our analyses using nonparametric Wilcoxon rank-sum tests that treat each group as one independent observation. These tests are the most conservative, because they disregard that we have observations from multiple participants and rounds per group. Nevertheless, the findings of these conservative tests are similar to those of the other analyses, indicating the robustness of our results. The Wilcoxon tests are provided in Supplementary Material section S6.

## 3. Results

### 3.1. Cooperation in multilevel public goods games

Fig. 2 shows the cooperation levels in multilevel and single-group PGGs for all four conditions. In this section, we focus on the results of the multilevel PGGs. In the conditions where the multilevel PGG comes after the single-group PGG (panels A and B in Fig. 2), global cooperation is higher than local cooperation. Although this difference between global cooperation and local cooperation is also present without punishment (8.31 MU vs. 4.94 MU, $p = .001$, Table S7), it is especially remarkable when punishment is possible (14.71 MU vs. 3.38 MU, $p < .001$, Table S7). Punishment strongly promotes global cooperation in these conditions (6.41 MU, $p < .001$, Table S3), and only slightly lowers local cooperation ($-1.56$ MU, $p = .04$, Table S2). Consequently, total cooperation is promoted by punishment (4.85 MU, $p < .001$, Table S1). This means that we find support for the hypothesized positive effect of punishment on total cooperation in multilevel PGGs that come after a single-group PGG.

The pattern is strikingly different in the conditions where the multilevel PGG comes before the single-group PGG (Fig. 2C and D). Here, we find a roughly equal mix of local and global cooperation. Punishment does not promote cooperation in these conditions; the levels of local and global cooperation are not significantly different between the non-punishment and punishment conditions (average local contributions of 7.20 MU vs. 8.43 MU, $p = .17$, Table S2; average global contributions of 8.25 MU vs. 8.45 MU, $p = .85$, Table S3; average total contributions of 15.45 MU vs. 16.88, $p = .05$, Table S1). This means that we do not find support for the hypothesized positive effect of punishment on total cooperation in multilevel PGGs that come before the single-group PGG.

**Result 1.** The hypothesis that cooperation in multilevel PGGs is higher when punishment is possible is only partially supported (H1). Punishment does not promote cooperation when people start with a multilevel PGG, but does promote cooperation when the multilevel PGG is preceded by a single-group PGG.

When punishment is not possible, we do not find a difference in global cooperation between the conditions where a single-group PGG comes before the multilevel PGG and conditions with the opposite order (8.31 MU vs 8.25 MU, $p = .96$, Table S7, and see Fig. 2A vs C). When punishment is possible, global cooperation is much higher in the condition where a single-group PGG comes before the multilevel PGG than in the condition with the opposite order (14.71 vs. 8.45 MU, $p < .001$, Table S7, and see Fig. 2B vs D). This runs counter to our hypothesis that global cooperation will be higher when the multilevel public goods problem is present from the start than when it arises after an initial single-group public goods problem.

**Result 2.** The hypothesis that global cooperation will be higher when the multilevel PGG is present from the start than when it comes after a single-group PGG is not supported (H2). In fact, global cooperation is higher when the multilevel PGG comes after a single-group PGG and punishment is possible.

With regard to time trends, we see that in conditions *multilevel to single-group* (with and without punishment, Fig. 2C and D and Tables S5 and S6), global contributions are being replaced by local contributions over time. Local contributions significantly increase over time (Table S5) while global contributions significantly decrease over time (Table S6). This means that contributions are increasingly separated into two local goods instead of a unified global good, indicating a trend of fragmentation. In the condition *single-group to multilevel* without punishment, local cooperation is stable over time (Table S5 and Fig. 2A) but global cooperation still decreases over time (Table S6 and Fig. 2A). Only in the condition *single-group to multilevel* with punishment is global cooperation stable

and high (Table S6 and Fig. 2B). As mentioned, the last round forms an exception to the time trends because norms were elicited just before this round. Norm elicitations can make norms temporarily more salient, thereby promoting cooperation in the last round (see Methods for details).

The norm elicitations could theoretically also be involved in our observed order effects because they are included before part 2 of the experiment. However, two pieces of evidence contradict this conjecture. First, before part 1 we also included some norm elicitations, namely those for personal normative views. Although personal normative views are not exactly the same as normative/empirical expectations (the additional norm elicitations before part 2), the aforementioned high correlations between them suggest they are highly similar for participants. This makes it unlikely that potential effects of norm elicitations before part 2 strongly differ from potential effects of norm elicitations before part 1 (and hence produce order effects). Second, we can compare the contributions directly before and after the norm elicitation (round 9 vs 10 and 19 vs 20) to obtain an indication of the effect of the norm elicitation. We examine the effect of norm elicitations in particular for conditions with punishment, because that is where we observed our order effects (i.e., cooperation differences between *single-group to multilevel* and *multilevel to single-group*). In single-group PGGs with punishment, local contributions are 0.99 MU higher directly after the norm elicitation ($t(111) = 2.32$, $p = .02$) and global contributions 0.04 MU lower ($t(111) = -0.40$, $p = .69$). In multilevel PGGs with punishment, local contributions are 0.80 MU lower directly after the norm elicitation ($t(111) = -1.55$, $p = .12$) and global contributions 0.85 MU higher ($t(111) = 1.35$, $p = .18$). These differences directly before and after the norm elicitation are relatively small and mostly statistically insignificant, suggesting they are unlikely to explain our large order effects.

### 3.2. Cooperation in single-group public goods games

In this section, we focus on the results of the single-group PGGs. As mentioned, the collectively efficient form of cooperation in single-group PGGs is local cooperation, and we see in Fig. 2 that virtually all cooperation in single-group PGGs is indeed local cooperation. This suggests that participants have little difficulty coordinating on the efficient way of cooperating in single-group PGGs. Consistent with previous research, we find that cooperation in single-group PGGs is higher when peer punishment is possible (3.40 MU higher in condition *single-group to multilevel,* $p < .001$, Fig. 2B vs A; 3.75 MU higher in condition *multilevel to single-group,* $p < .001$, Fig. 2D vs C; see Supplementary Material Table S1).

Further in line with previous research, we find that there is a decaying trend in cooperation without punishment (Fig. 2A and C, Supplementary Material Table S4). Note that, again, we find an exception to these trends in the last round because norms were elicited just before this round (see Methods for details). If punishment is possible, cooperation is sustained at a high level throughout all rounds (Fig. 2B and D). A *t*-test comparing the 3.40 MU increase in contributions due to punishment in condition *single-group to multilevel* with the 3.75 MU increase in condition *multilevel to single-group* finds that the increase is independent of order ($t(2199) = 0.28$, $p = .78$). In sum, punishment promotes cooperation in single-group PGGs regardless of whether the single-group PGG comes before or after the multilevel PGG. In addition to the time series of Fig. 2, we provide in Table 1 the summary statistics for the contributions and punishment averaged over the ten rounds by condition and PGG type.

**Table 1**
Descriptive statistics on contributions and punishment by condition and PGG.

| | | With punishment | | Without punishment | |
|---|---|---|---|---|---|
| | | Single-group to multilevel | Multilevel to single-group | Single-group to multilevel | Multilevel to single-group |
| Single-group PGGs | Local contributions | 16.90 | 17.94 | 13.10 | 13.97 |
| | | (5.60) | (4.52) | (7.23) | (7.72) |
| | Global contributions | 0.60 | 0.48 | 1.00 | 0.70 |
| | | (2.44) | (1.50) | (2.81) | (2.88) |
| | Total contributions | 17.50 | 18.42 | 14.10 | 14.67 |
| | | (5.01) | (3.62) | (6.98) | (7.45) |
| | Punishment points | 0.45 | 0.32 | – | – |
| | | (1.70) | (1.06) | – | – |
| Multilevel PGGs | Local contributions | 3.38 | 8.43 | 4.94 | 7.20 |
| | | (5.13) | (6.12) | (5.37) | (6.62) |
| | Global contributions | 14.71 | 8.45 | 8.31 | 8.25 |
| | | (6.83) | (6.94) | (6.78) | (7.22) |
| | Total contributions | 18.09 | 16.88 | 13.24 | 15.45 |
| | | (4.20) | (4.37) | (6.22) | (6.05) |
| | Punishment points | 0.59 | 1.01 | – | – |
| | | (1.84) | (1.78) | – | – |

Note: Values are means over the ten rounds by condition and public good game (single-group and multilevel). Standard deviations are given in parentheses. Punishment points refer to received punishment points. Contributions could be between 0 and 20 MU and punishment points between 0 and 10 per co-member.

### 3.3. Punishment patterns

Recall that we suggested that punishment may be less effective in promoting cooperation in multilevel public goods problems than in single-group public goods problems because the potential for normative disagreement is higher in multilevel public goods problems. Prior research suggests that normative disagreement is reflected in more punishment (Nikiforakis et al., 2012; Rauhut and Winter, 2017; Winter et al., 2012, 2018). Hence, if normative disagreement is higher in multilevel public goods problems, we would expect more punishment in multilevel PGGs. Indeed, we find that punishment occurs more often in multilevel PGGs than in single-group PGGs (participants in multilevel PGGs punish at least one group member 29.5 % of the time compared to 16.0 % in single-group PGGs, $p <$

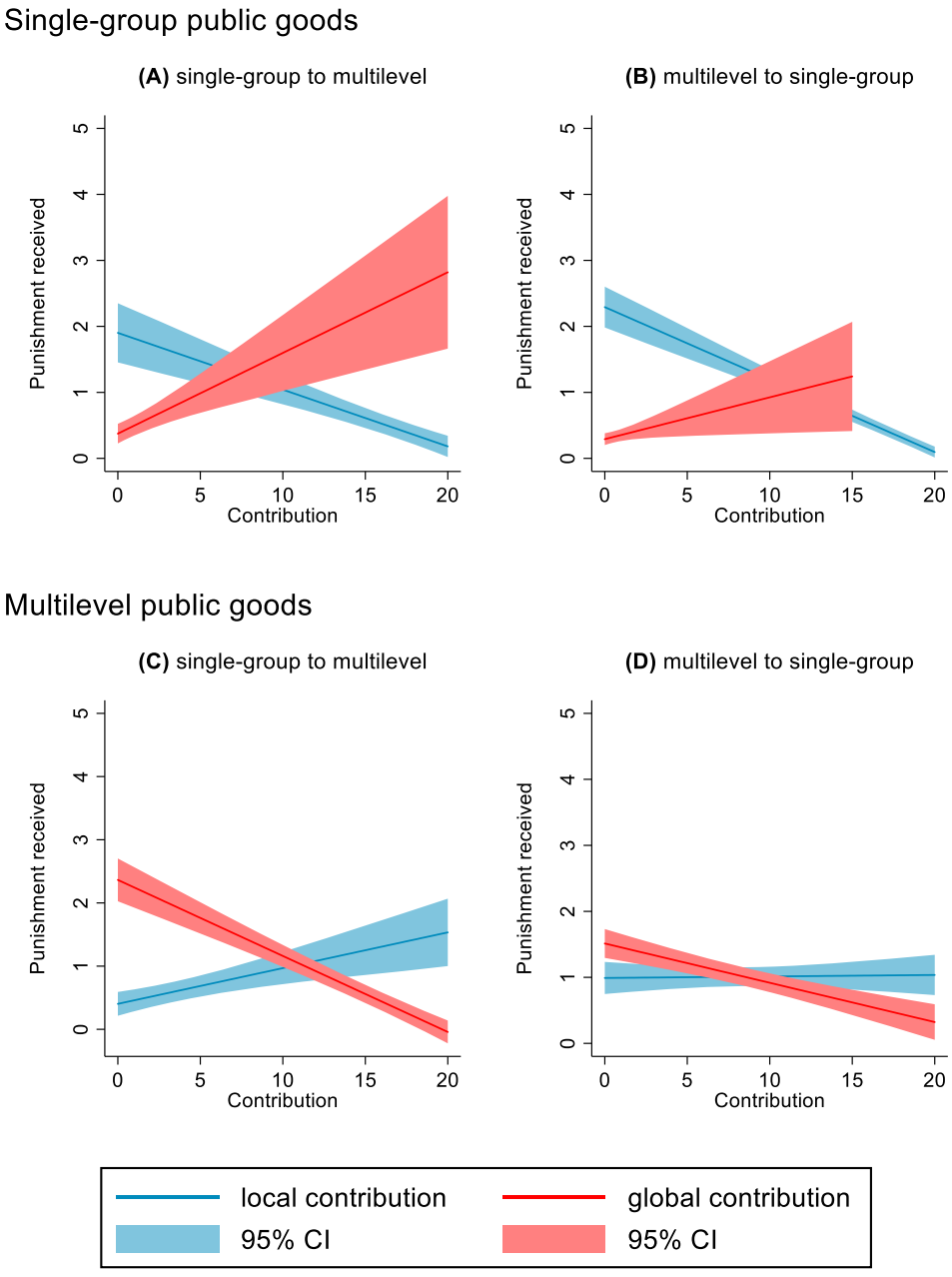



**Fig. 3.** Punishment patterns in single-group and multilevel public goods games. We present the linear relationships between received punishment and contribution behavior, and 95 % confidence intervals are included via shaded areas. Received punishment refers to the total number of punishment points received in a round by a player; each punishment point reduces the punished player's payoff by 3 MU. In single-group public goods games (panels A and B), we find that punishment increases with lower local contributions and higher global contributions. In multilevel public goods games (panels C and D), we find that punishment increases with lower global contributions and higher local contributions, in particular in the condition where a single-group public goods game comes before the multilevel public goods game (panel C).

.001, Table S8). However, this difference is only significant in the condition where the multilevel PGG comes before the single-group PGG (36.8 % vs 16.3 %, $p < .001$, Table S8). In the condition where a single-group PGG comes before the multilevel PGG, punishment is not significantly higher in the multilevel PGG than in the single-group PGG (21.0 % vs 15.6 %, $p = .19$, Table S8). This suggests that initial experience in single-group PGGs may help to prevent normative disagreement in subsequent multilevel PGGs and reduce the extent of punishment necessary to maintain cooperation.

Recall further that we expected groups that start with a single-group PGG to develop norms of local contributions, which then potentially carry over to subsequent multilevel PGGs and thereby lead to local instead of global contributions. Groups that start with a multilevel PGG are not constrained by these norms of local contributions that developed in the single-group PGG and may therefore more easily achieve global contributions. Because punishment is regarded as a means of norm enforcement, we can examine which contribution behaviors are punished to get a first indication of which norms are present in single-group and multilevel PGGs and whether these depend on the order of the problem. In Fig. 3, we show what type of contribution behavior is punished in single-group and multilevel PGGs. In single-group PGGs (Fig. 3A and B), punishment is directed at low contributions to local goods and high contributions to global goods. Participants contributing nothing to local goods receive 2.11 punishment points whereas those contributing fully to local goods receive 0.14 punishment points (the difference is 1.97 punishment points, $p < .001$, Table S10). Participants contributing nothing to global goods receive 0.33 punishment points whereas those contributing fully to global goods receive 2.14 punishment points (the difference is 1.81 punishment points, $p < .001$, Table S10). These punishment patterns do not differ significantly between the condition where the single-group PGG comes before the multilevel PGG (Figure A) and the condition with the opposite order (Fig. 3B; Table S9). The punishment patterns are in line with a norm of contributing to local goods in single-group PGGs.

In multilevel PGGs (Fig. 3C and D), punishment is directed at high contributions to local goods and low contributions to global goods, but mostly in the condition where a single-group PGG comes before the multilevel PGG (Fig. 3C; Table S9). In this condition, participants contributing nothing to local goods receive 0.40 punishment points whereas those contributing fully to local goods receive 1.53 punishment points (the difference is 1.13 punishment points, $p < .001$, Table S10). Participants contributing nothing to global goods receive 2.36 punishment points whereas those contributing fully to global goods receive no punishment points (the difference is

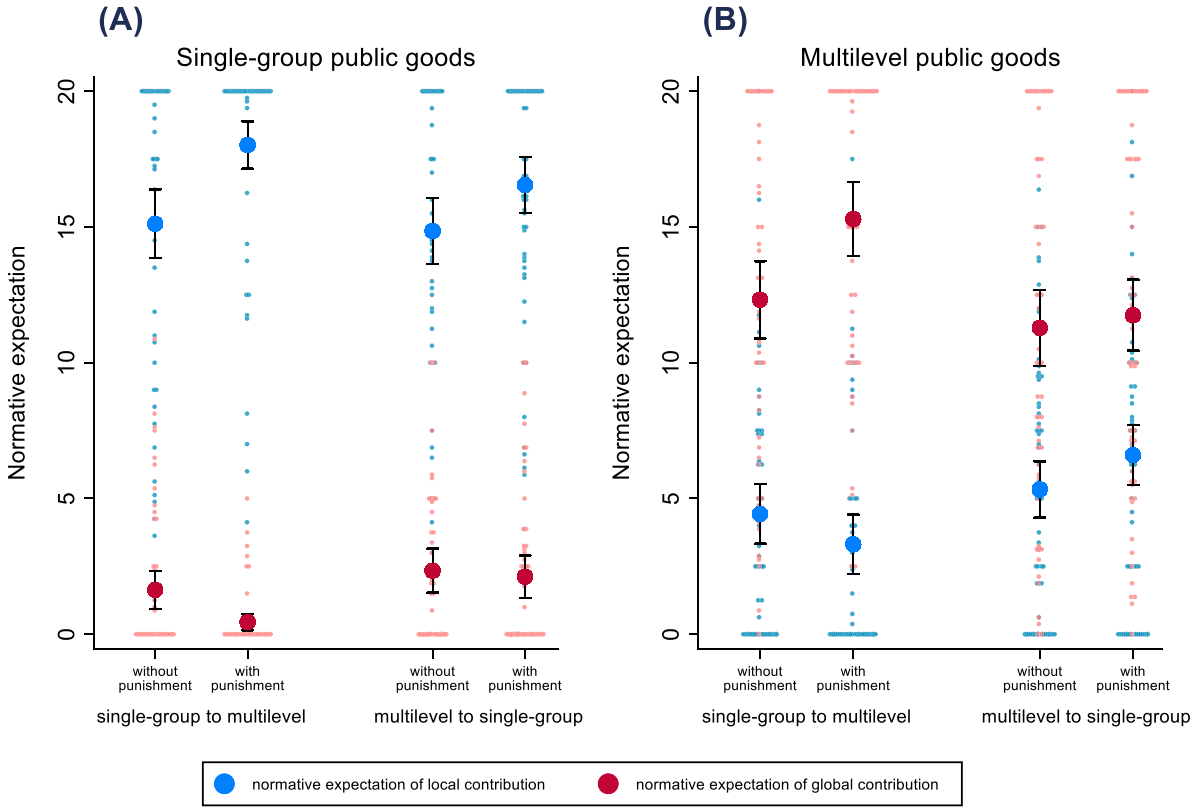



**Fig. 4.** Norms in single-group and multilevel public goods games. We present the normative expectations averaged over the first and second sets of 10 rounds. The results separated by measurement moment are available in the supplementary material Figures S2 and S3. The large markers show mean values and the small markers show the values of individual participants. The 95 % confidence intervals are included via capped spikes. In single-group public goods games (A), we find that normative expectations of local cooperation are higher if peer punishment is possible. In multilevel public goods games (B), we find that normative expectations of global cooperation are higher if peer punishment is possible, but only in the condition where participants interacted in a single-group public goods game before the multilevel public goods game.

significant at $p < .001$, Table S10). In the condition with the opposite order (Fig. 3D), punishment is relatively independent of how people contribute, but a tendency for punishment of low contributions to global goods is also discernable. In this condition, contributions to local goods are not significantly related to received punishment (null-contributors receive 0.99 punishment points compared to 1.04 punishment points for full contributors, $p = .85$, Table S10). Contributing to global goods is still associated with lower received punishment, but the association is significantly smaller than in the condition where a single-group PGG comes before the multilevel PGG (Table S9). These punishment patterns are in line with a norm of contributing to global goods in multilevel PGGs, in particular when a single-group PGG was encountered before the multilevel PGG. Hence, in contrast to our expectations, groups with prior experience in single-group PGGs seem to more easily achieve a norm of global cooperation in multilevel PGGs than groups that start with the multilevel PGG.

In Supplementary Material Figure S1, we show how people respond to received punishment depending on their contribution behavior. In single-group PGGs, low local contributors typically respond with higher local contributions, in particular in the condition where a multilevel PGG comes before the single-group PGG. High local contributors typically do not change their behavior when being punished, and neither do the few low or high global contributors (global contributions in single-group PGGs may reflect a lack of understanding the game, which may also explain why global contributors do not respond to punishment). In multilevel PGGs, reactions to punishment strongly depend on whether a single-group PGG comes before the multilevel PGG or vice versa. In the condition where a single-group PGG comes before the multilevel PGG, low global contributors react with higher global contributions, high local contributors react with lower local contributions, and high global contributors or low local contributors do not react to punishment. These reactions are in line with a norm of global cooperation in multilevel PGGs. In the condition where the multilevel PGG comes before the single-group PGG, low global contributors also react with higher global contributions, but low and high local contributors do not react to punishment, and high global contributors respond to punishment with lower global contributions. These reactions are less clearly in line with the norm of global cooperation.

### 3.4. Norms in single-group and multilevel public goods games

Results on the contribution and punishment behavior of the participants suggest a norm of local contributions in single-group PGGs and a norm of global contributions in multilevel PGGs when a single-group PGG comes before the multilevel PGG. We now examine whether the norms suggested by these contribution and punishment patterns are reflected in the direct measurements of norms. To measure norms, we asked participants to report their normative expectations, i.e., what they expect their group members think are appropriate contributions to the local and global account in both PGGs (single-group and multilevel). The results are shown in Fig. 4. We see that for single-group PGGs, participants hold normative expectations of high local contributions. These normative expectations are significantly higher with punishment (2.90 MU higher in condition *single-group to multilevel,* $p = .003$; 1.69 MU higher in condition *multilevel to single-group,* $p = .06$; Table S11), suggesting that punishment may not only act as a means of norm enforcement but can also increase the perception of the norm.

For multilevel public goods, we see that normative expectations are higher regarding global contributions than local contributions in all four conditions. This norm of global contributions is strongest in the condition where a single-group PGG comes before the multilevel PGG and where punishment is possible. That is, the normative expectations of global contributions in this condition are significantly higher than in the other three conditions (2.97 MU higher than in *single-group to multilevel* without punishment, $p = .02$; 3.55 MU higher than in *multilevel to single-group* with punishment, $p = .003$; 4.01 MU higher than in *multilevel to single-group* without punishment, $p = .001$; Table S12). Recall that groups in condition *single-group to multilevel* with punishment also achieved the highest actual levels of global contributions (Fig. 2) and most clearly punished deviations from global contributions (Fig. 3). Groups in the other three conditions also started with relatively high global contributions, but showed a decline in global contributions over time. These lower levels of global contributions are consistent with these three conditions having lower norms of global contributions (Fig. 4). Altogether, the normative expectations resemble the observed cooperation and punishment patterns rather closely for both the single-group and multilevel PGGs.

We had hypothesized that global cooperation would be lower in multilevel PGGs when participants have prior experience with single-group PGGs. This unsupported hypothesis was based on the underlying assumption that individuals' experience in single-group PGGs will lead to norms of local cooperation, which then will carry over to the multilevel PGG. Our norm elicitations allow us to examine whether this underlying assumption holds. In particular, because we measured norms in round 10 of part 1 about both single-group and multilevel PGGs regardless of which game participants played themselves, we can test the underlying assumption that local norms in single-group PGGs carry over to local norms in multilevel PGGs. Using the findings from the norm elicitation at round 10 of part 1 (see Supplementary Material Figure S2), we can see that the underlying assumption did indeed not hold. Experience with single-group PGGs leads to normative expectations of local cooperation for single-group PGGs but normative expectations of global cooperation for multilevel PGGs. This indicates an overarching norm of collective efficiency instead of a norm of local cooperation for both types of PGGs.

However, we can still examine whether the norm that developed during the single-group PGG in part 1 is then influential in part 2 when participants actually play the multilevel PGG. That is, given that a norm of collective efficiency developed in condition *single-group to multilevel* in part 1, does this norm carry over to part 2? Our findings suggest it does; participants contribute largely according to this norm in part 2 (Fig. 2B), and their normative expectations at part 1 correlate with their contributions in part 2. The correlation between normative expectations of local contribution in round 10 of part 1 and participants' actual local contributions in the subsequent ten rounds of part 2 is 0.58 ($p < .001$); for global contributions the correlation is 0.55 ($p < .001$). Hence, there is still evidence for norm stickiness, but based on a different norm than we had expected would develop. Note also that this means that the

(small) subset of participants who did develop normative expectations of local cooperation for multilevel PGGs while first playing the single-group PGG, also contributed more locally in the subsequent multilevel PGG than others who developed normative expectations of global cooperation.

Another underlying assumption was that there is more normative disagreement in multilevel PGGs than in single-group PGGs. To assess this assumption, we examine the overlap in normative expectations between group members. Normative disagreement is lowest when all group members hold the same normative expectations and is highest when group members all hold different normative expectations. In Supplementary Material Figure S4, we show the extent of overlap in normative expectations between group members for single-group and multilevel PGGs in all four conditions. We indeed find that normative disagreement is generally higher in multilevel PGGs than in single-group PGGs. The only exception is when a single-group PGG comes before the multilevel PGG and punishment is possible. Participants in this condition have a relatively high overlap in normative expectations between members even in multilevel PGGs. This again suggests that initial experience with solving single-group PGGs may help to prevent normative disagreement in subsequent multilevel PGGs.

## 4. Discussion

A large body of research shows that punishment promotes cooperation in single-group public goods problems (Balliet et al., 2011; Chaudhuri, 2011; Fehr and Gächter, 2002; Fehr and Schurtenberger, 2018; Henrich et al., 2006). Because several real-life public goods problems involve multiple group memberships, we studied whether punishment also promotes cooperation in multilevel public goods problems. We furthermore examined how the effect of punishment depends on whether groups are in a single-group public goods problem before the multilevel public goods problem or vice versa. Results show that while punishment promotes cooperation in single-group public goods problems, punishment does not promote cooperation in multilevel public goods problems if groups do not have prior experience in single-group public goods problems. In these groups, there is a roughly equal mix of local and global cooperation and there is a decay in global cooperation over time even with punishment. Hence, we observe some fragmentation into separate local goods at the expense of the more collectively beneficial global good. However, in groups that do have prior experience with single-group public goods problems, punishment does promote and sustain cooperation in the multilevel public goods problem. What is more, punishment leads these groups to develop and enforce collectively efficient norms of global cooperation and thus prevents fragmented clusters of local cooperation.

Our findings establish important boundary conditions for the effects of punishment on cooperation. Whereas we found punishment to promote cooperation in single-group public goods problems regardless of whether this problem appeared before or after the multilevel public goods problem, the effect of punishment in multilevel public goods problems does crucially depend on this order. That punishment does not promote cooperation in groups that start with multilevel public goods problems challenges the view that punishment unequivocally promotes cooperation. Moreover, the observed punishment patterns and norm measurements suggest that normative disagreement is generally higher in multilevel public goods problems than in single-group public goods problems. However, for groups that moved from a single-group public goods problem to a multilevel public goods problem, normative disagreement was not higher in the multilevel problem. What is more, contrary to expectations, experience with single-group public goods problems did not lead to norms of local cooperation in subsequent multilevel public goods problems. If anything, initial experience with solving single-group public goods problems helped to develop and enforce norms of global cooperation in subsequent multilevel public goods problems.

Our findings suggest that 'starting small'– by moving from single-group public goods problems to multilevel public goods problems – is a promising strategy for achieving global cooperation in multilevel public goods problems. One may wonder whether the initial experience with single-group public goods problems can be replaced by more experience with multilevel problems. That is, perhaps people would also cooperate more in a second multilevel PGG after experience with a first multilevel PGG due to a restart effect. However, the behavioral patterns observed in our experiment are considerably different from typical restart effects. When comparing the first-round global contribution of groups with and without prior experience in single-group PGGs, we find it to be 50 % higher in groups with the prior experience. This difference further increases after the first round (74 % over all ten rounds). In contrast, restart effects typically show people in the second game contributing similarly to how they contributed at the start of the first game (rarely noticeably higher), and then contributions decrease again (Andreoni, 1988; Croson, 1996; Masclet et al., 2003). Hence, the patterns that we observe are considerably different from patterns involved in restart effects, both in terms of magnitude and developments over time. Although a restart effect could partly be involved in high global contributions after the single-group PGG, it is unlikely to fully explain the high level of global contributions. This suggests that it is not only experience with the multilevel problem in general that helps to achieve cooperation, but also the 'starting small' with a single-group public goods problem before turning to the multilevel public goods problem.

This result complements previous research showing that 'starting small' helps to achieve cooperation in public good provision. For example, experiments show that achieving cooperation in large groups is facilitated by slowly increasing the group size over time instead of immediately starting with a large group size (Charness and Yang, 2014; Salmon and Weber, 2017; Weber, 2006), and that achieving cooperation in high-stake situations is facilitated by slowly increasing the stakes over time instead of starting immediately with high stakes (Ye et al., 2020). Moving from smaller-scale economies to larger-scale economies has also been shown to promote more efficient systems of exchange and specialisation than starting with larger-scale economies (Crockett et al., 2009). More generally, starting with simple versions of a problem before turning to more complex versions of a problem has been shown to help task performance (Yasarcan, 2009). Our findings are further in line with recent research suggesting that 'local-to-global' mechanisms help to achieve cooperation for global problems (Hauser et al., 2016; McGinnis and Ostrom, 2008). Finally, the results are in line with the idea

of an 'expanding circle of cooperation'; people first cooperate within small units such as families and subgroups before being able to cooperate in larger social units involving other subgroups and strangers (Singer, 2011; Smith, 2017).

Multilevel structures in society often arise over time from membership changes between local groups. A well-known example is the increasingly multicultural structure of contemporary western societies, in which individuals from different ethnic groups increasingly need to cooperate. A common conjecture is that interaction in local groups leads to norms of cooperation specifically suited to the current group members (Bernhard et al., 2006; Choi et al., 2019, 2021; Titlestad et al., 2019). When these groups end up in a multilevel problem through mixing with other groups, the preexisting norms of local cooperation may lead to fragmented clusters of local cooperation and impede global cooperation across local groups. Our findings suggest this need not always be the case. Initial interaction in local groups can also facilitate a norm of cooperation that maximizes efficiency, and groups realize that global cooperation maximizes efficiency when they end up in a multilevel problem through group changes. However, the attainment of global cooperation does crucially depend on the option to enforce norms via peer punishment. When punishment is not possible, we observe a downward trend in global cooperation, also in groups with prior experience in single-group public goods problems.

Because we are the first to test the effect of punishment in multilevel public goods experiments, we stayed relatively close to the typical paradigm of public goods games with peer punishment. We offer a few suggestions for future research to test whether our findings generalize to other situations. First, the local groups in our multilevel public goods problem were of equal size and had equal punishing power to enforce norms. These conditions are not always realized in real-life groups or societies, where the incumbent group is usually the majority and has more power to impose its norms on the incoming members from different groups (Otten et al., 2021). In future experiments, these conditions can be reproduced by making one of the local groups larger in size than the other (Otten et al., 2021) and/or by alternating who can punish whom (Ozono et al., 2020).

Because experience in single-group PGGs did not lead to norms of local cooperation for multilevel PGGs, we could not fully test the stickiness of norms of local cooperation. Although we did find evidence for norm stickiness in the norms that did develop, the norm stickiness of local cooperation in particular remains to be further studied. To do so, future research could employ experimental designs that lead to norms of local cooperation even for multilevel PGGs, which then allows to test norm stickiness for local cooperation in particular. One way is to introduce contexts that induce more animosity between local groups, for example involving intergroup conflict for scarce resources. Another way is to move from minimal groups to natural groups with pre-existing animosities from real life (Drouvelis et al., 2021; Weisel and Böhm, 2015).

Our norm elicitations allowed us to discover that different norms developed than initially assumed, but including norm elicitations does carry the risk of affecting behavior, e.g., by making norms temporarily more salient. What is more, previous evidence suggests that making norms salient may improve the effectiveness of punishment somewhat (Andrighetto et al., 2013; Bicchieri et al., 2021). Our observed punishment effects may therefore have been weaker without the norm elicitations, although we doubt that the effect of punishment in multilevel PGGs with prior experience in single-group PGGs would disappear given the large effect size. A previous study finds little evidence for order effects due to norm elicitations (D'Adda et al., 2016), but future research to disentangle the effects of punishment and norm elicitations may be worthwhile.

Future research could also further examine differences between single-group and multilevel PGGs, for example in terms of reputation formation. Reputation formation may be easier when there exist several small local groups within a bigger global group (multilevel public goods) than when everybody is part of one big local group (single-group public goods). This was also the case in our experiment, where participants could more easily identify the 1 other local member in the multilevel PGG than each of the 3 other local members in the single-group PGG. The effect of these differences in reputation formation can go both ways. On the one hand, the larger reputation formation possibilities in the local group of the multilevel PGG could lead to more efficient (global) cooperation, because reputation has been shown to promote efficient cooperation (Bolton et al., 2005). On the other hand, when reputations are limited to one's own local group, they may form an incentive to act in the interest of that local group, leading to local instead of global contributions. Delving into these reputation formation processes could therefore help to further understand the cooperation differences between single-group PGGs and multilevel PGGs.

We found that punishment only promotes cooperation in multilevel public goods problems when groups have prior experience with solving single-group public goods problems. In an increasingly interconnected world, multiple group memberships may exacerbate the challenges of achieving collective cooperation. Different local groups may pursue different goals, potentially leading to fragmented clusters of local cooperation that prevent the joint benefits from global cooperation. Our results suggest that 'starting small' in combination with opportunities for norm enforcement may help to overcome such challenges.

## Funding

Netherlands Organization for Scientific Research (NWO) Gravitation Grant 024.003.025 and funding from the Hebrew University and Utrecht University Partners of Science (HUPS).

## Declarations of competing interest

None.

## Acknowledgments

This study is part of the research program Sustainable Cooperation – Roadmaps to Resilient Societies (SCOOP). The authors are

grateful to the Netherlands Organization for Scientific Research (NWO) and the Dutch Ministry of Education, Culture and Science (OCW) for generously funding this research in the context of its 2017 Gravitation Program (grant number 024.003.025). We also thank the Hebrew University and Utrecht University Partners of Science (HUPS) for funding the data collection.

## Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.euroecorev.2024.104682.

## References

Aaldering, H., Böhm, R., 2020. Parochial versus universal cooperation: introducing a novel economic game of within- and between-group interaction. Soc. Psychol. Personal Sci. 11 (1), 36–45. https://doi.org/10.1177/1948550619841627.
Aaldering, H., Ten Velden, F.S., Van Kleef, G.A., De Dreu, C.K.W., 2018. Parochial cooperation in nested intergroup dilemmas is reduced when it harms out-groups. J. Pers. Soc. Psychol. 114 (6), 909–923. https://doi.org/10.1037/pspi0000125.
Andreoni, J., 1988. Why free ride? Strategies and learning in public goods experiments. J. Public Econ. 37 (3), 291–304. https://doi.org/10.1016/0047-2727(88)90043-6.
Andreoni, J., Nikiforakis, N., Siegenthaler, S., 2021. Predicting social tipping and norm change in controlled experiments. Proc. Natl. Acad. Sci. 118 (16) https://doi.org/10.1073/pnas.2014893118.
Andrighetto, G., Brandts, J., Conte, R., Sabater-Mir, J., Solaz, H., Villatoro, D., 2013. Punish and voice: punishment enhances cooperation when combined with norm-signalling. PLoS One 8 (6), 1–8. https://doi.org/10.1371/journal.pone.0064941.
Balafoutas, L., Nikiforakis, N., 2012. Norm enforcement in the city: a natural field experiment. Eur. Econ. Rev. 56 (8), 1773–1785. https://doi.org/10.1016/j.euroecorev.2012.09.008.
Balafoutas, L., Nikiforakis, N., Rockenbach, B., 2014. Direct and indirect punishment among strangers in the field. Proc. Natl. Acad. Sci. 111 (45), 15924–15927. https://doi.org/10.1073/pnas.1413170111.
Balliet, D., Mulder, L.B., Van Lange, P.A.M., 2011. Reward, punishment, and cooperation: a meta-analysis. Psychol. Bull. 137 (4), 594–615. https://doi.org/10.1037/a0023489.
Balliet, D., Wu, J., De Dreu, C.K.W., 2014. Ingroup favoritism in cooperation: a meta-analysis. Psychol. Bull. 140 (6), 1556–1581. https://doi.org/10.1037/a0037737.
Bernhard, H., Fischbacher, U., Fehr, E., 2006. Parochial altruism in humans. Nature 442 (7105), 912–915. https://doi.org/10.1038/nature04981.
Bicchieri, C., Chavez, A., 2010. Behaving as expected: public information and fairness norms. J. Behav. Decis. Mak. 23, 161–178. https://doi.org/10.1002/bdm.648.
Bicchieri, C., Dimant, E., Xiao, E., 2021. Deviant or wrong? The effects of norm information on the efficacy of punishment. J. Econ. Behav. Organ. 188, 209–235. https://doi.org/10.1016/j.jebo.2021.04.002.
Blackwell, C., McKee, M., 2003. Only for my own neighborhood? Preferences and voluntary provision of local and global public goods. J. Econ. Behav. Organ. 52 (1), 115–131. https://doi.org/10.1016/S0167-2681(02)00178-6.
Böhm, R., Bornstein, G., Koppel, H., 2014. Between-group conflict and other-regarding preferences in nested social dilemmas. Jena Econ. Res. Pap.
Bolton, G.E., Katok, E., Ockenfels, A., 2005. Cooperation among strangers with limited information about reputation. J. Public Econ. 89 (8 SPEC. ISS), 1457–1468. https://doi.org/10.1016/j.jpubeco.2004.03.008.
Bowles, S., Gintis, H., 2013. A Cooperative Species: Human Reciprocity and Its Evolution. Princeton University Press.
Buchan, N.R., Brewer, M.B., Grimalda, G., Wilson, R.K., Fatas, E., Foddy, M., 2011. Global social identity and global cooperation. Psychol. Sci. 22 (6), 821–828. https://doi.org/10.1177/0956797611409590.
Buchan, N.R., Grimalda, G., Wilson, R., Brewer, M., Fatas, E., Foddy, M., 2009. Globalization and human cooperation. Proc. Natl. Acad. Sci. 106 (11), 4138–4142. https://doi.org/10.1073/pnas.0809522106.
Catola, M., Alessandro, S.D., Guarnieri, P., Pizziol, V., & Alessandro, S.D., 2021. Personal and social norms in a multilevel public goods experiment (Working Paper).
Charness, G., Yang, C., 2014. Starting small toward voluntary formation of efficient large groups in public goods provision. J. Econ. Behav. Organ. 102, 119–132. https://doi.org/10.1016/j.jebo.2014.03.005.
Chaudhuri, A., 2011. Sustaining cooperation in laboratory public goods experiments: a selective survey of the literature. Exp. Econ. 14 (1), 47–83. https://doi.org/10.1007/s10683-010-9257-1.
Choi, D.D., Poertner, M., Sambanis, N., 2019. Parochialism, social norms, and discrimination against immigrants. Proc. Natl. Acad. Sci. 116 (33), 16274–16279. https://doi.org/10.1073/pnas.1820146116.
Choi, D.D., Poertner, M., Sambanis, N., 2021. The hijab penalty: feminist backlash to Muslim immigrants. Am. J. Pol. Sci. 00 (0), 1–16. https://doi.org/10.1111/ajps.12627.
Cotta, B., 2020. What goes around, comes around? Access and allocation problems in Global North–South waste trade. Int. Environ. Agreem.: Polit., Law Econ. 20 (2), 255–269. https://doi.org/10.1007/s10784-020-09479-3.
Crockett, S., Smith, V.L., Wilson, B.J., 2009. Exchange and specialisation as a discovery process. Econ. J. 119 (539), 1162–1188. https://doi.org/10.1111/j.1468-0297.2009.02254.x.
Croson, R.T.A., 1996. Partners and strangers revisited. Econ. Lett. 53, 25–32. https://doi.org/10.1016/S0165-1765(97)82136-2.
Cubitt, R.P., Drouvelis, M., Gächter, S., Kabalin, R., 2011. Moral judgments in social dilemmas: how bad is free riding? J. Public Econ. 95 (3–4), 253–264. https://doi.org/10.1016/j.jpubeco.2010.10.011.
D'Adda, G., Drouvelis, M., Nosenzo, D., 2016. Norm elicitation in within-subject designs: testing for order effects. J. Behav. Exp. Econ. 62, 1–7. https://doi.org/10.1016/j.socec.2016.02.003.
Degen, D., Kuhn, T., Van der Brug, W., 2019. Granting immigrants access to social benefits? How self-interest influences support for welfare state restrictiveness. J. Eur. Soc. Policy 29 (2), 148–165. https://doi.org/10.1177/0958928718781293.
Dimant, E., 2022. Hate trumps love: the impact of political polarization on social preferences. Manage. Sci. https://doi.org/10.2139/ssrn.3848335.
Drouvelis, M., Malaeb, B., Vlassopoulos, M., Wahba, J., 2021. Cooperation in a fragmented society: experimental evidence on Syrian refugees and natives in Lebanon. J. Econ. Behav. Organ. 187, 176–191. https://doi.org/10.1016/j.jebo.2021.04.032.
Duffy, J., Lafky, J., 2021. Social conformity under evolving private preferences. Games Econ. Behav. 128, 104–124. https://doi.org/10.1016/j.geb.2021.04.005.
Efferson, C., Vogt, S., 2018. Behavioural homogenization with spillovers in a normative domain. Proc. R. Soc. B: Biol. Sci. 285 (1879) https://doi.org/10.1098/rspb.2018.0492.
Engl, F., Riedl, A.M., Weber, R.A., 2021. Spillover effects of institutions on cooperative behavior, preferences, and beliefs. Am. Econ. J.: Microecon. 13 (4), 261–299. https://doi.org/10.1257/mic.20180336, 2021.
Espinosa, M.P., Fatás, E., Ubeda, P., 2019. Linguistic diversity and out-group discrimination in bilingual societies. J. Behav. Exp. Econ. 81, 102–127. https://doi.org/10.1016/j.socec.2019.06.002. July 2018.
Fehr, E., Gächter, S., 2000. Cooperation and punishment in public goods experiments. Am. Econ. Rev. 90 (4), 980–994. https://doi.org/10.1257/aer.90.4.980.
Fehr, E., Gächter, S., 2002. Altruistic punishment in humans. Nature 415 (6868), 137–140. https://doi.org/10.1038/415137a.

Fehr, E., Schurtenberger, I., 2018. Normative foundations of human cooperation. Nat. Hum. Behav. 2 (7), 458–468. https://doi.org/10.1038/s41562-018-0385-5.
Fellner, G., Lünser, G.K., 2014. Cooperation in local and global groups. J. Econ. Behav. Organ. 108, 364–373. https://doi.org/10.1016/j.jebo.2014.02.007.
Fischbacher, U., 2007. Z-Tree: Zurich toolbox for ready-made economic experiments. Exp. Econ. 10 (2), 171–178. https://doi.org/10.1007/s10683-006-9159-4.
Gächter, S., Renner, E., Sefton, M., 2008. The long-run benefits of punishment. Science 322 (5907), 1510. https://doi.org/10.1126/science.1164744.
Gallier, C., Goeschl, T., Kesternich, M., Lohse, J., Reif, C., Römer, D., 2019. Leveling up? An inter-neighborhood experiment on parochialism and the efficiency of multi-level public goods provision. J. Econ. Behav. Organ. 164, 500–517. https://doi.org/10.1016/j.jebo.2019.05.028.
Greiner, B., 2015. Subject pool recruitment procedures: organizing experiments with ORSEE. J. Econ. Sci. Assoc. 1 (1), 114–125. https://doi.org/10.1007/s40881-015-0004-4.
Grimson, D., Knowles, S., Stahlmann-Brown, P., 2020. How close to home does charity begin? Appl. Econ. 52 (34), 3700–3708. https://doi.org/10.1080/00036846.2020.1720906.
Guala, F., Mittone, L., 2010. How history and convention create norms: an experimental study. J. Econ. Psychol. 31 (4), 749–756. https://doi.org/10.1016/j.joep.2010.05.009.
Hauser, O.P., Hendriks, A., Rand, D.G., Nowak, M.A., 2016. Think global, act local: preserving the global commons. Sci. Rep. 6, 1–7. https://doi.org/10.1038/srep36079.
Hawkes, C., Altman, J., Beckerman, S., Grinker, R.R., Harpending, H., Jeske, J., Peterson, N., Smith, E.A., Wenzel, G.W., Yellen, J.E., 1993. Why hunter gatherers work: an ancient version of the problem of public goods. Curr. Anthropol. 34 (4), 341–361. https://doi.org/10.1086/204182.
Henrich, J., McElreath, R., Barr, A., Ensminger, J., Barrett, C., Bolyanatz, A., Cardaroas, J.C., Gurven, M., Gwako, E., Henrich, N., Lesoronol, C., Marlowe, F., Tracer, D., Ziker, J., 2006. Costly punishment across human societies. Science 312 (5781), 1767–1770. https://doi.org/10.1126/science.1127333.
Herrmann, B., Thöni, C., Gächter, S., 2008. Antisocial punishment across societies. Science 319 (5868), 1362–1367. https://doi.org/10.1126/science.1153808.
Hill, K., 2002. Altruistic cooperation during foraging by the Ache, and the evolved human predisposition to cooperate. Hum. Nat. 13 (1), 105–128. https://doi.org/10.1007/s12110-002-1016-3.
Israel, S., Weisel, O., Ebstein, R.P., Bornstein, G., 2012. Oxytocin, but not vasopressin, increases both parochial and universal altruism. Psychoneuroendocrinology 37 (8), 1341–1344. https://doi.org/10.1016/j.psyneuen.2012.02.001.
Kallhoff, A., 2014. Why societies need public goods. Crit. Rev. Int. Soc. Polit. Philos. 17 (6), 635–651.
Kimbrough, E.O., Vostroknutov, A., 2016. Norms make preferences social. J. Eur. Econ. Assoc. 14 (3), 608–638. https://doi.org/10.1111/jeea.12152.
Krupka, E.L., Weber, R.A., 2009. The focusing and informational effects of norms on pro-social behavior. J. Econ. Psychol. 30 (3), 307–320. https://doi.org/10.1016/j.joep.2008.11.005.
Krupka, E.L., Weber, R.A., 2013. Identifying social norms using coordination games: why does dictator game sharing vary? J. Eur. Econ. Assoc. 11 (3), 495–524. https://doi.org/10.1111/jeea.12006.
Lange, A., Schmitz, J., Schwirplies, C., 2022. Inequality, role reversal and cooperation in multiple group membership settings. Exp. Econ. 25 (1), 68–110. https://doi.org/10.1007/s10683-021-09705-y.
Masclet, D., Noussair, C., Tucker, S., Villeval, M.C., 2003. Monetary and nonmonetary punishment in the voluntary contributions mechanism. Am. Econ. Rev. 93 (1), 366–380. https://doi.org/10.1257/000282803321455359.
McGinnis, M., Ostrom, E., 2008. Will lessons from small-scale social dilemmas scale up? In: Biel, A., Eek, D., Gärling, T., Gustafsson, M. (Eds.), New Issues and Paradigms in Research On Social Dilemmas. Springer, pp. 189–211.
Nikiforakis, N., Noussair, C.N., Wilkening, T., 2012. Normative conflict and feuds: the limits of self-enforcement. J. Public Econ. 96 (9–10), 797–807. https://doi.org/10.1016/j.jpubeco.2012.05.014.
Olson, M., 1965. The Logic of Collective Action: Public Goods and the Theory of Groups. Harvard University Press.
Ostrom, E., 2000. Collective action and the evolution of social norms. J. Econ. Perspect. 14 (3), 137–158. https://doi.org/10.1257/jep.14.3.137.
Otten, K., Buskens, V., Przepiorka, W., Ellemers, N., 2020. Heterogeneous groups cooperate in public good problems despite normative disagreements about individual contribution levels. Sci. Rep. 16702, 1–12. https://doi.org/10.1038/s41598-020-73314-7.
Otten, K., Buskens, V., Przepiorka, W., Ellemers, N., 2021. Cooperation between newcomers and incumbents: the role of normative disagreements. J. Econ. Psychol. 87, 102448 https://doi.org/10.1016/j.joep.2021.102448.
Otten, K., Frey, U.J., Buskens, V., Przepiorka, W., Ellemers, N., 2022. Human cooperation in changing groups in a large-scale public goods game. Nat. Commun. 13 (1), 1–11. https://doi.org/10.1038/s41467-022-34160-5.
Ozono, H., Kamijo, Y., Shimizu, K., 2020. The role of peer reward and punishment for public goods problems in a localized society. Sci. Rep. 10 (1), 1–6. https://doi.org/10.1038/s41598-020-64930-4.
Peysakhovich, A., Rand, D.G., 2013. Habits of virtue: creating norms of cooperation and defection in the laboratory. Manage. Sci. 62 (3), 631–647. https://doi.org/10.1021/la00030a006.
Polzer, J.T., Milton, L.P., & Gruenfeld, D.H., 2009. Asymmetric subgroup communication in nested social dilemmas (Working Paper).
Polzer, J.T., Stewart, K.J., Simmons, J.L., 1999. A social categorization explanation for framing effects in nested social dilemmas. Organ Behav. Hum. Decis. Process. 79 (2), 154–178. https://doi.org/10.1006/obhd.1999.2842.
Przepiorka, W., Szekely, A., Andrighetto, G., Diekmann, A., Tummolini, L., 2022. How norms emerge from conventions (and change). Socius 8. https://doi.org/10.1177/23780231221124556.
Rauhut, H., Winter, F., 2017. Types of normative conflicts and the effectiveness of punishment. In: Przepiorka, W., Jann, B. (Eds.), Social Dilemmas, Institutions, and the Evolution of Cooperation, pp. 239–256.
Reuben, E., Riedl, A., 2013. Enforcement of contribution norms in public good games with heterogeneous populations. Games Econ. Behav. 77 (1), 122–137. https://doi.org/10.1016/j.geb.2012.10.001.
Reuben, E., Riedl, A., & Bernard, M., 2015. Fairness and coordination: the role of fairness principles in coordination failure and success (Working Paper). https://sites.google.com/site/markbernard1984/research-papers.
Romano, A., Spadaro, G., Balliet, D., Joireman, J., Lissa, C.Van, Jin, S., Agostini, M., Bélanger, J.J., Gützkow, B., Kreienkamp, J., Leander, N.P., Abakoumkin, G., Khaiyom, J.H.A., Ahmedi, V., Akkas, H., Almenara, C.A., Atta, M., Bagci, S.C., Basel, S., Zúñiga, C., 2021. Cooperation and trust across societies during the COVID-19 pandemic. J. Cross Cult. Psychol. 52 (7), 622–642. https://doi.org/10.1177/0022022120988913.
Salmon, T.C., Weber, R.A., 2017. Maintaining efficiency while integrating entrants from lower performing groups: an experimental study. Econ. J. 127 (600), 417–444. https://doi.org/10.1111/ecoj.12308.
Schultz, C., 1996. Polarization and inefficient policies. Rev. Econ. Stud. 63 (2), 331–344. https://doi.org/10.2307/2297855.
Singer, P., 2011. The Expanding Circle: Ethics, Evolution, and Moral Progress. Princeton University Press.
Smerdon, D., Offerman, T., Gneezy, U., 2019. 'Everybody's doing it': on the persistence of bad social norms. Exp. Econ., 0123456789 https://doi.org/10.1007/s10683-019-09616-z.
Smith, S.R., 2017. Modelling "the expanding circle" of cooperation towards a sustainable future. Eur. J. Sustain. Dev. 6 (4), 341–352. https://doi.org/10.14207/ejsd.2017.v6n4p341.
Spadaro, G., Tiddi, I., Columbus, S., Jin, S., Teije, A.ten, Team, C., Balliet, D., 2022. The Cooperation Databank: machine-readable science accelerates research synthesis. Perspect. Psychol. Sci. 1–18. https://doi.org/10.1177/17456916211053319.
Tavoni, A., Dannenberg, A., Kallis, G., Loschel, A., 2011. Inequality, communication, and the avoidance of disastrous climate change in a public goods game. Proc. Natl. Acade. Sci. 108 (29), 11825–11829. https://doi.org/10.1073/pnas.1102493108.
Titlestad, K., Snijders, T.A.B., Durrheim, K., Quayle, M., Postmes, T., 2019. The dynamic emergence of cooperative norms in a social dilemma. J. Exp. Soc. Psychol. 84 (103799) https://doi.org/10.1016/j.jesp.2019.03.010.
Van Dijk, E., De Dreu, C.K.W., 2021. Experimental games and social decision making. Annu. Rev. Psychol. 72, 415–438. https://doi.org/10.1146/annurev-psych-081420-110718.

Weber, R., 2006. Managing growth to achieve efficient coordination in large groups: theory and experimental evidence. Am. Econ. Rev. 96 (1), 114–126. https://doi.org/10.1257/000282806776157588.
Weber, R.A., Camerer, C.F., 2003. Cultural conflict and merger failure: an experimental approach. Manage. Sci. 49 (4), 400–415. https://doi.org/10.1287/mnsc.49.4.400.14430.
Weisel, O., Böhm, R., 2015. Ingroup love" and "outgroup hate" in intergroup conflict between natural groups. J. Exp. Soc. Psychol. 60, 110–120. https://doi.org/10.1016/j.jesp.2015.04.008.
Winter, F., Rauhut, H., Helbing, D., 2012. How norms can generate conflict: an experiment on the failure of cooperative micro-motives on the macro-level. Soc. Forces 90 (3), 919–948. https://doi.org/10.1093/sf/sor028.
Winter, F., Rauhut, H., Miller, L., 2018. Dynamic bargaining and normative conflict. J. Behav. Exp. Econ. 74, 112–126. https://doi.org/10.1016/j.socec.2018.03.003.
Wit, A.P., Kerr, N.L., 2002. Me versus just us versus us all" categorization and cooperation in nested social dilemmas. J. Pers. Soc. Psychol. 83 (3), 616–637. https://doi.org/10.1037/0022-3514.83.3.616.
Yasarcan, H., 2009. Improving understanding, learning, and performances of novices in dynamic managerial simulation games. Complexity 15 (4), 31–42. https://doi.org/10.1002/cplx.20292.
Ye, M., Zheng, J., Nikolov, P., Asher, S., 2020. One step at a time: does gradualism build coordination? Manage. Sci. 66 (1), 113–129.
Ye, Y., Zhang, Q., Wei, X., Cao, Z., Yuan, H.Y., Zeng, D.D., 2022. Equitable access to COVID-19 vaccines makes a life-saving difference to all countries. Nat. Hum. Behav. 6 (2), 207–216. https://doi.org/10.1038/s41562-022-01289-8.
Young, H., 2015. The evolution of social norms. Annu. Rev. Econom. 7 (1), 359–387. https://doi.org/10.1146/annurev-economics-080614-115322.