ISSN 1755-5361

University of Essex

Department of Economics

Discussion Paper Series

No. 750 March 2014

Exclude the bad actors or learn about the group

David Hugh-Jones and David Reinstein

Exclude the bad actors or learn about the group

David Hugh-Jones and David Reinstein*

March 20, 2014

Abstract

In public goods environments, the threat to punish non-contributors may increase contributions. However, this threat may make players' contributions less informative about their true social preferences. This lack of information may lead to lower contributions after the threat disappears, as we show in a two stage model with selfish and conditionally cooperative types. Under specified conditions welfare may be improved by committing not to punish or exclude. Our laboratory evidence supports this. Contributions under the threat of targeted punishment were less informative of subjects' later choices than contributions made anonymously. Subjects also *realized* that these were less informative, and their incentivized predictions reflected this understanding. We find evidence of conditional cooperation driven by beliefs over others' contributions. Overall, our Anonymous treatment led to lower first-stage contributions but significantly higher second-stage contributions than our Revealed treatment. Our model and evidence may help explain why anonymous contributions are often encouraged in the real world.

Keywords: signaling, anonymity, public goods, club goods, experiments, social trust, reciprocity

Authors: David Hugh-Jones, University of Essex, Wivenhoe Park, Colchester CO4 3SQ, United Kingdom; dhughj@essex.ac.uk. David Reinstein (corresponding author), University of Essex, Wivenhoe Park, Colchester CO4 3SQ, United Kingdom; drein@essex.ac.uk; +44 1206 87 3518.

JEL codes: H41, Z12, D82.

*David Hugh-Jones is a lecturer in Government at the University of Essex. David Reinstein is a lecturer in Economics, also at the University of Essex. We thank Toru Suzuki, Ondrej Rydval, Gerlinde Fellner, Martin Leroch, Ryan Mackay, Vittoria Levatti and Henry Bottomley ; seminar participants at the University of Queensland, Sungkyunkwan University, the University of Oxford, the Max Plank Institute, Hamburg University, the University of Essex, the University of Nottingham, the University of Warwick, the University of Amsterdam, and M-BEES. Huajing He, Jonathan Homola, Yousef Makhlouf, Liutauras Petrucionis, provided excellent research assistance.

1 Introduction

There is evidence that punishment and exclusion can increase cooperation in laboratory public goods games. Yet, in real world public goods situations, we do not always observe punishment or exclusion. Indeed, sometimes contribution mechanisms seem to be structured to avoid the possibility of direct punishment - for example, by making contributions anonymous. In this paper, we provide and model an explanation, and present a laboratory experiment offering supporting evidence. We show that punishment and exclusion can prevent players learning about each others' true preferences, and lower mutual trust. In some cases, making punishment less targeted can then make institutions more efficient.

Consider a group whose members can benefit from each other's cooperation, such as farmers who must work together to bring in a harvest, workplace colleagues who can produce more by exerting non-contractible effort, union members who can support each other in disputes with management, or a unit of soldiers during war. Each of these situations involve a collective action problem: individually rational, self-interested actors will "defect", leading to an inefficient outcome. Cooperation in such groups may be on-going; if the interactions are repeated indefinitely, then even self-interested actors may cooperate (Fudenberg and Maskin, 1986). However, even within a long-term relationship there may occur *extreme episodes*, periods in which immediate payoffs or losses are larger then normal, offering a greater temptation to behave selfishly (cf. Rotemberg et al., 1986 in an oligopoly context). For example, farmers may experience poor harvests, a firm may be threatened with bankruptcy, a union may begin a prolonged strike, and soldiers may face frontline combat. In these *extreme episodes* the present benefits of selfishness may outweigh the future cost of a breakdown of cooperation, and previously cooperative individuals may defect. Similarly, cooperation may be sustained by outside enforcement, but this enforcement may break down during anarchic periods.

Some individuals (e.g., those with lower discount rates or a stronger other-regarding preference) may cooperate during an extreme episode or anarchic period. These individuals may be *conditional cooperators* who prefer to cooperate only if others do so as well. There is strong evidence that some, but not all, individuals are conditional cooperators.[1]

We model the situation just described as two rounds of a public goods game. The first stage represents an everyday interaction. The second stage, which has a greater impact on total welfare, represents an extreme episode.[2] In the second stage, self-interested players will always defect, but conditional cooperators will cooperate only if they expect others to do so too: i.e., if they believe that enough other players are conditional cooperators, and if they expect those players to cooperate.[3]

[1] See Ledyard (1993), Isaac, R. M., J. Walker, and S. Thomas (1984), Ostrom (2000), Plott and Smith (2008) section 6.1, and Chaudhuri (2011) for surveys; we discuss only the tip of the iceberg below.

[2] This simple structure captures the signaling logic of an infinitely repeated game with randomly occurring extreme episodes. It is also analogous to a finitely repeated game with some "critical" stage, past which self-interested players no longer cooperate, following the logic of Kreps et al. (2001).

[3] We motivate this with a vignette. Describing a strike on a Midwestern university campus, Dixon and Roscigno (2003)

There are three important connections between the everyday interaction and the extreme episode, i.e., between the first and second stages. First, whenever the public good has some rivalry the group may benefit if selfish types are excluded from the second stage. For example, unproductive farmers may burden the whole community during a famine, and union members who work during a strike will undermine a union's negotiating power. As a result, if exclusion is possible (i.e., for a "club good"), selfish types may be excluded.[4]

Second, the group may benefit if the number of selfish types becomes common knowledge. Knowing that there are many selfish types may reduce cooperation by conditional cooperators, while knowing there are many cooperative types will have the opposite effect. Better information may help on average, depending on the ex-ante beliefs and on the payoff functions. For example, if conditional cooperators are ex-ante pessimistic and would cooperate little under their *uninformed prior* beliefs, then information that leads to a positive surprise may increase their contributions dramatically, while disappointing information will have little effect. . "Where information helps" (described formally in Lemma 1), the knowledge of the group's type gained from the first-stage interaction will improve average second-stage welfare.

Lastly, all players, including selfish ones, have an incentive to act like conditional cooperators during the first stage, since by doing so they may convince others there is "one more good type", and thus spur second stage cooperation. This incentive may be small, since the resulting increase in others' cooperation is itself a public good. However, the incentive to pool will become much stronger when selfish behavior in the first stage *also* brings the risk of being excluded.

This last consequence creates a conflict, from the point of view of group welfare, between the potential benefit of learning the number of selfish types and the incentive to exclude selfish types. If selfish types face exclusion or punishment then they are likely to mimic the behavior of conditional cooperators to avoid this, and this makes it harder for others to learn of their presence. Yet, the *ex post* incentives to exclude the selfish from the second stage may be strong enough that promises not to do so will not be credible.

In this context, anonymity can serve as a group self-commitment device. The everyday interaction may be structured so that the profile of contributions is visible, but no individual contributor can be identified. (For example, in a labor union, dues may be collected anonymously but reported in sum, and elections can use a secret ballot but report the level of turnout.) As a result, individually targeted

quote a union activist: "while workers were signing up for picket duty in the week leading up to the strike, and certainly talking to one another and 'sizing each other up,' many made up their minds at the last possible moment." Strike participation was highly correlated within work units. This suggests that workers had conditionally cooperative preferences and updated their beliefs about overall participation based on the behavior of their work unit peers.

[4]Even without crowding, punishment or exclusion may be driven by psychological motives and moral norms, such as fairness and reciprocity (Fehr and Schmidt, 2006). Indeed, exclusion or punishment of the unproductive is often observed. For example, workers may be "sent to Coventry" (socially ostracized) by their union colleagues if they are seen to behave in a way that hurts other workers, such as putting in "excessive" effort, failing to contribute union dues or to show up to meetings. In a religious context, many US protestant churches practice "shunning"; similar practices include *cherem* in Judaism, disfellowshipping among Jehovah's witnesses and disconnection in Scientology.

exclusion or punishment becomes impossible. Selfish types then have weaker incentives to act like cooperative types in the first stage. Contributions in the first stage will better reflect the players' types, and therefore will be more informative of play in the second stage. Conditional cooperators then learn more, and, *where information helps*, the level of second-stage cooperation will increase on average.

In section 3 we present a model with a small Voluntary Contributions Mechanism (VCM) game followed by an exclusion decision, followed by a larger VCM.[5] While all players get the same linear "material" payoff from the total contribution, *conditionally cooperative* types also get a payoff that is complementary in one's own contribution and others'.[6] Types are private information; in Lemma 1 we give conditions under which *learning* others' types increases contributions in the VCM. We next consider the exclusion choice, and the equilibrium in the full game, parametrizing the relative size of the first and second VCM. Even if exclusion cannot be targeted based on first-stage contributions, selfish types will pool with conditional cooperators (to send a false signal of "one more cooperator" and raise second-stage contributions) unless the first-stage VCM is sufficiently large. However, when exclusion can be targeted, implying contributions (more strongly) reduce the probability of being excluded, the first VCM must be made even *larger* to separate the types, as we show in Proposition 2. Putting together Lemma 1 and Proposition 2, our model predicts that less targeted enforcement will lead to greater revelation of true types, and that in net this *may* increase efficiency in public goods settings.[7]

To test the plausibility of our theory, we run a laboratory experiment. In each of 15 repetitions subjects played two sequential public goods games, with an exclusion decision after stage 1, and re-assignment to a new group after stage 2. In our *Revealed* treatment subjects could target someone for exclusion based on her stage 1 contribution; in our *Anonymous* treatment subjects learned only the aggregate profile of contributions and could choose only to exclude someone at random. As our model predicted, *Revealed* stage 1 contributions were significantly less informative signals of subjects' stage 2 contributions than *Anonymous* stage 1 contributions. Other subjects' incentivized predictions reflected this. We find evidence of conditional cooperation driven by *beliefs* over others' stage-2 contribution. Our *Anonymous* treatment also led to lower stage 1 contributions but significantly higher stage 2 contributions.

This suggests a partial explanation for the preservation of anonymity in some public goods environments, which is puzzling in light of evidence that anonymity decreases contributions (e.g. Harbaugh, 1998; Glazer and Konrad, 1996; Milinski et al., 2002; Cooter and Broughman, 2005; Soetevent, 2005;

[5] We refer to these games as *VCMs* following the standard terminology in the experimental literature. In our context these are essentially impure public goods, involving both some exclusion and some rivalry or "crowding".

[6] The latter payoff may also be material; e.g., some types may benefit more financially from *contributing* to a succesful effort. However, our experiment depends on conditional cooperators having a primal psychological motivation.

[7] Hugh-Jones and Reinstein (2012) formalize a "burning money" variation of this theory. In the present paper the signaling game takes the same form as the main public goods game. It seems natural that the signaling institution might resemble the basic collective action problem; as good types benefit more from their own contribution, it becomes cheaper for them to signal (as in the standard model of Spence, 1973), and thus easier to separate the types.

Andreoni and Petrie, 2004; Alpizar et al., 2008). As noted above, anonymity may serve as a commitment to less targeted punishment, as contributions cannot be linked back to any one player, reducing the incentive for selfish types to pool with conditionally cooperative types. Thus, even if anonymity reduces contributions in everyday interactions, it may allow a better measure of a group's underlying type, build mutual trust, and thus better coordinate later cooperation when enforcement is impossible.

This suggests a hidden benefit of several institutions which encourage anonymous contributions, including church donations, religious norms of private giving, the secret ballot, and group incentive schemes. Many religions encourage anonymous contributions; Matthew 6:2-4 enjoins "But when thou doest alms, let not thy left hand know what thy right hand doeth..." Although this may not be the most effective way of raising money, the anonymity may be critical to building mutual trust within religious communities.

Voting is often anonymous. One reason is to prevent private actors from bribing or threatening voters, but the anonymity may also lead to a more honest signal; in particular, as there is some cost to voting, a high turnout may signal the strength of public concern. In the 1980's, the UK Conservative government forced unions to hold secret ballots before a strike, expecting this to reduce the effectiveness of strike threats. In fact, the ballot may have given the unions leverage to negotiate with management (Martin et al., 1991). In workplaces where not supporting the union might result in social disapproval, precisely the *anonymity* of the ballot box made it a credible signal of members' willingness to support industrial action. Lastly, the management literature on team-building discourages "finger-pointing"; truly cohesive teams must fail or succeed collectively (Katzenbach et al., 2001). We discuss anonymous institutions at greater length in Hugh-Jones and Reinstein (2012).

The remainder of our paper is organized as follows. In section 2 we discuss the related theoretical and experimental literature. Section 3 presents our model. We discuss experimental design issues in 4.1, describe our specific design choices in 4.2, and give summary statistics in 4.3. We derive our experimental hypotheses in section 4.4, and we test these and present related results in section 4.5. We conclude in Section 5 with an interpretation and further motivation for our results, and suggestions for future work.

2 Literature

A variety of papers have discussed and modeled the signaling value of public goods contributions (Veblen, 1899; Glazer and Konrad, 1996; Katz and Rosenberg, 2005; Carpenter and Myers, 2010) but these are all about signaling an *individual's own* trustworthiness. Our model more closely resembles that of Londregan and Vindigni (2006), where an individual's signal provides information about the *group*.

2.1

In most VCM and public goods experiments, where the self-interested dominant strategy is to contribute nothing,[8] contributions decline over repetitions (among the same group), but remain substantially greater than zero. There is considerable evidence for heterogeneous preferences: some subjects are self-interested, others are conditional cooperators. In explaining the standard patterm, Ostrom (2000) argues that conditional cooperators (good types) begin optimistic, and some selfish types strategically pool with them (as in Kreps et al., 2001), but this gradually unravels. On the other hand, when conditionally cooperative types are isolated, or can separate themselves, a higher level of cooperation can be sustained. As Ostrom (2000) puts it, "a core question is how potential cooperators signal one another and design institutions that reinforce rather than destroy conditional cooperation". This may occur naturally, as subjects observe each others' play over repeated interactions. However, as our model suggests, the temptation to to punish or exclude free-riders may lead to pooling behavior and impede signaling and learning. Thus monitoring and punishment may lead to worse outcomes if and when these institutions are no longer effective.[9]

A wide class of papers investigate the impact of the threat of punishment or exclusion (related to "ostracism") in VCM's. Most of these papers (e.g., Fehr and Gächter, 2000, Cinyabuguma et al., 2005) find that these threats increase contributions while the threats are present. However, the net effect on efficiency is often ambiguous (see, e.g., Bochet et al., 2006) under costly punishment. Furthermore, Herrmann et al. (2008) document how "antisocial punishment" (the punishment of high contributors) occurs in certain groups, and this can make the punishment regime counterproductive. We could find no experimental evidence on the impact of such punishment institutions on *later* cooperation after punishment is no longer available or feasible (e.g., see literature surveys by Ledyard, 1993 and Chaudhuri, 2011).

Several recent experiments (e.g., Coricelli et al., 2004) examine the implications of heterogeneity through classifying and sorting subjects by "type." Burlando and Guala (2005) use pre-game measures to classify players' types, and then put subjects into homogeneous groups to play the public goods game. Contributions remain high among the more cooperative groups, and overall contributions are increased by the segregation. However, these papers do not inform the subjects in advance what these tests will be used for, and therefore rule out strategic misrepresentation of preferences. In the real world, such classification mechanisms are prone to manipulation.

Other papers (beginning with Erhart and Keser, 1999) investigate *endogenous* group formation. Ahn et al. (2009) allow restricted entry and exit (by majority vote) in various treatments. Here group

[8] See Ledyard (1993), Isaac, R. M., J. Walker, and S. Thomas (1984), Ostrom (2000), Plott and Smith (2008) section 6.1, and Chaudhuri (2011) for surveys; we discuss only the tip of the iceberg below.

[9] Broadly speaking the idea that weakening the monitoring may lead to better information echoes Ichino and Muehlheusser (2008). However, their work involves the monitoring of an individual agent and does not involve conditional cooperation and the signaling of the "average group type," as in our model.

size is endogenous, and behavior may be strategic; in their restricted entry treatment players may build a history of contribution to get admitted to a cooperative group. Indeed, in the restricted entry treatment they observe a significant "endgame effect," with sharp declines in cooperation. Coricelli et al. (2004) and Cinyabuguma et al. (2005) find similar endgame effects, as do Keser and van Winden (2000) particularly in their "partners" treatment.[10]The evidence is mixed, but it appears that there are limits to endogenous sorting – it does not guarantee that conditional cooperators will be able to identify and associate with each other.

Our experiment explores this further. Since our groups are re-matched between repetitions of the two-stage game, there is an "endgame" in every second stage. Thus we are able to measure the effect of targeted exclusion on (i) contributions under the threat of exclusion, (ii) contributions after this threat is removed, and (iii) beliefs about other players' contributions.

3 Model

To fix ideas and demonstrate internal consistency we represent our theory formally. All proofs are in Appendix A.1. Each of $i = 1...N$ players choose to contribute $x_i \geq 0$ to a collective good. Let $\bar{X} = \sum_{j=1}^{N} x_j/N$ be average contributions and $\bar{X}_{-i} = \sum_{j\neq i} x_j/(N-1)$ be the average contribution of all players except player i. There are two types of players. Player i's welfare is given by

$$\alpha\bar{X} - x_i + w(x_i, \bar{X}_{-i}) \tag{1}$$

if she is a conditionally cooperative, or "good" type, and

$$\alpha\bar{X} - x_i \tag{2}$$

if she is a selfish or "bad" type. Player types are independent: the probability of a good type is $\pi \in (0,1)$. Let $\Pi(g)$ represent any player's expectation that there are exactly g good types among the remaining $N-1$ players; this comes from a binomial distribution. $\alpha \in (1,N)$ is the multiplier for the unconditional benefit of the collective good (we will refer to $\alpha\bar{X} - x_i$ as the "material payoff"). We denote the marginal unconditional benefit of one's own contributions as $\overline{\alpha} = \alpha/N$. In addition, w represents a benefit received by good types only, which is twice differentiable, strictly concave and strictly increasing in both its arguments, with a positive cross partial, and $w(0,0) = 0$; we will refer to this payoff as the "CC payoff" (for "conditional cooperator").

To ensure the existence of an interior equilibrium, we make the following assumptions**.** First, good types always prefer to contribute something, i.e., $w_1(0,0) > 1 - \bar{\alpha}$. Second, the marginal return to contributions ultimately diminishes towards zero, implying that everyone's preferred contribution is

[10]In contrast Page et al. (2005) find much smaller end-game effects for their endogenous "regrouping" treatments. However, their "final stage" only occurs once, and represents a small share of expected payoffs; this limits players' incentive and ability to learn strategic contribution behavior.

bounded. In other words, the derivatives w_1 and w_2 are bounded, with $w_1(x_i, \bar{X}_{-i}) \to 0$ as $x_i \to \infty$ for all $\bar{X}_{-i}$; and there is some finite $\widetilde{x}$ with $w_1(\widetilde{x}, \bar{X}_{-i}) < 1 - \bar{\alpha}$ for all $\bar{X}_{-i}$. To guarantee a unique equilibrium, we also make the technical assumption that $\frac{w_{12}(x,X)}{w_{11}(x,X)} \in (-k\frac{x}{X}, 0)$ for all x, all $X > 0$ and some fixed $k \in (0,1)$.

By construction, a bad player never contributes anything, since $\bar{\alpha} < 1$. A good player who expects that others' average contributions will be exactly X solves the first order condition and contributes x_i such that

$$w_1(x_i, X) = 1 - \bar{\alpha}. \tag{3}$$

Define the best response as $b(X) \equiv x_i$ satisfying the above. This is single-valued by the strict concavity of w.

When others' contributions are uncertain, good types' optimal contributions satisfy

$$E_{\bar{X}_{-i}} w_1(x_i, \bar{X}_{-i}) = 1 - \bar{\alpha}. \tag{4}$$

Suppose that, prior to the game, there are revealed to be exactly $g+1$ good players in total (with the identity of these players either common knowledge, or completely unknown). Then there is a unique equilibrium in which each good type contributes the same amount $x_g > 0$. On the other hand, if players only know the "prior" distribution of good players, this again implies a unique symmetric equilibrium, in which all good types contribute $x^* > 0$.

Comparing these two cases, common knowledge of types will increase contributions on average when

$$\sum_{g=0}^{N-1} \Pi(g)(g+1)x^* < \sum_{g=0}^{N-1} \Pi(g)(g+1)x_g. \tag{5}$$

Here $(g+1)x^*$ or $(g+1)x_g$ gives the total value of equilibrium contributions when there are $g+1$ good types, which occurs with probability $\Pi(g)$.[11] This inequality need not hold in general; the following is a sufficient condition.

Lemma 1. *Common knowledge of the number of good types increases contributions ex ante (on average) when $w_1(x,X)$ is weakly concave in X and the best response function $b(\cdot)$ is weakly convex.*

The weak convexity of $b(\cdot)$ implies that a good type's optimal contribution increases in response to others' contributions at an increasing rate. As an example, the conditions of the above Lemma hold if the CC payoff is "Cobb-Douglas," i.e., if $w(x,X) = x^\gamma X^{1-\gamma}$ with $\gamma \in (0,1)$. Note that the weak convexity of $b(\cdot)$ is simply a condition on the locus of points (x,X) such that $w_1(x,X) = 1 - \bar{\alpha}$. Hence this is equivalent to a condition on the model's primals.

[11] This is assuming there is at least one good type. If there are no good types, contributions are zero in both cases.

The first stage

We assume that the first stage is simply a scaled-down version of the main collective action problem, in which both material and CC payoffs are multiplied by $D < 1$.[12] In between the stages, players may be excluded from the second round, in which case they receive zero utility from it. Thus, i's total welfare, using superscripts for stages 1 and 2, is

$$D\left[\alpha\bar{X}^1 - x_i^1 + \tau_i w(x_i^1, \bar{X}_{-i}^1)\right] + I_i\left[\alpha\bar{X}^2 - x_i^2 + \tau_i w(x_i^2, \bar{X}_{-i}^2)\right]$$

where $\tau_i = 1$ if i is good and 0 otherwise, and $I_i = 1$ if player i is included in the second stage, $I_i = 0$ otherwise.

There are two different types of first stage. After a *revealed* first stage all players' contributions are public knowledge. After an *anonymous* first stage only the profile of contributions is revealed, so exclusion cannot be targeted at any particular player based on her round 1 contribution.

For technical simplicity, we assume that exclusions are implemented only when they will increase the proportion of good types in the second round. Thus, in the revealed institution, if the first round behavior is distinct for each type**,** only good types are included in the second round, while if there is pooling in the first round, there will be no exclusion. In the anonymous institution, where bad individuals cannot be targeted for exclusion, no players are excluded. However, Proposition 2 below will hold even if we allow a certain proportion to be excluded under anonymity, as long as the slope of the probability of being excluded in one's contribution is shallower under anonymity.

We look for a separating equilibrium, where the types play differently in the first stage. When D is very small, the incentive to pretend to be good and thus increase round 2 contributions will dominate the incentive to contribute little in the first stage, even for bad type players, and separation will be impossible. When D is larger, a separating equilibrium will be possible. An anonymous first stage allows separating equilibria for lower values of D. The reason is intuitive: when play is anonymous, the cost of playing selfishly is only that others contribute less in the second round. When play is revealed, selfish play results in exclusion from the second round. As a result, the incentive for bad types to pool with good types, and to contribute in round 1, is greater in the revealed institution.

We refine our set of equilibria using the Intuitive Criterion. Here, this requires that in equilibrium, good types cannot profitably deviate towards the contribution that they prefer in the stage-game, unless a bad type can also profit from making such a deviation.

Proposition 2. *In the anonymous first stage, there is an Intuitive separating equilibrium (only) for values of D above a fixed value $\hat{D}$. In the revealed first stage, there is an Intuitive separating equilibrium (only) for values of D above a fixed value D^*, where $D^* > \hat{D}$.*

[12] Other assumptions are possible: for example, first round material welfare could be the same but with smaller maximum contributions, $x_i \in [0, D]$, with the CC payoff function unchanged as $w(x, X)$. Our formulation is chosen for simplicity, but our results are not sensitive to this assumption, since our proofs do not use the fact that w is identical between rounds.

If $D \in (\hat{D}, D^*)$, a separating equilibrium is possible in an anonymous first stage but not in a revealed first stage. In these separating equilibria, good types contribute in the first round, while bad types do not. In other words, the smallest first round that allows a separating equilibrium is larger when the first round is revealed than when it is anonymous. For an intermediate-sized first round, the types *only* separate under anonymity. Thus, under some conditions, especially when it is costly or impossible to increase the size of the first round, an anonymous first round may be preferred. This proposition suggests that an anonymous first stage may be preferred if common knowledge of the number of good types increases contributions on average (as in Lemma 1), and if the first stage is very small, or it is expensive to shift resources into the first stage.

As we demonstrate in section 4.4, our experimental setup seems to fall in the "intermediate" range: first-round play is more informative of second round play when the first round is anonymous, and this additional information seems to build trust and increase second stage cooperation.

4 Experiment

4.1 Design issues

In standard linear payoff VCM games, as in our experiment, with common knowledge of material self-interest, the standard prediction is zero contributions. However, if subjects are heterogeneous and have social preferences such as fairness and altruism, economic theory offers no clear prediction. As described above, previous experiments offer evidence for many of the *components* of our theory. We bring these together in a single context that measures subjects' inherent preferences and beliefs and observes the strategic responses to these and the resulting outcomes. Our model is rich: it makes several falsifiable predictions, specified below; our experiment is carefully designed to be able to test each of these, as cleanly as possible, in a unified setting. We find evidence that is consistent with each of our several hypotheses and inconsistent with reasonable alternative hypotheses.

Our experiment focuses on exclusion rather than costly punishment, as this seems more relevant to the real-world institutions we are considering. As noted below, exclusion amounts to a form of punishment in this context. We might have more simply compared exclusion to "no exclusion"; we include an exclusion decision in *both* treatments for parallelism. The prospect of being excluded, or of voting on the exclusion of others, may in itself affect contributions; we aim to balance this across treatments.[13]

Our hypotheses are formulated by *analogy* with our formal model, echoing the "instrumentalist" view of Friedman (1953) and the "fictionalist" approach in which a model connects "with the real world by relations of similarity" (Sitzia and Sugden, 2011). Our model is derived from the simplest assumptions and used to generate meaningful testable predictions for a more complex environment. Our

[13] In Appendix A.2 4.4 we offer evidence suggesting that our results are not driven by a differential probability of exclusion by treatment.

experimental design incorporates homegrown preferences and beliefs, and specific elements meant to rule out alternative explanations. Thus our experiment does not allow simple formally-derived predictions without making heroic assumptions. In essence, our design sacrifices a direct theoretical prediction to allow for more robust and powerful testing, in a more informative setting.[14]

4.2 Design and implementation

Our design is as follows. Thirty subjects enter the session; 15 are randomly assigned to the *Anonymous*, and 15 to the *Revealed* treatment. Subjects read through the instructions (shown in the online Appendix).[15] They then play 15 repetitions of a two-stage public goods game, always remaining within the same enforcement treatment, i.e., this is a between-subjects design.[16] For each repetition, subjects are randomly rematched into groups of five.[17] Each repetition is meant to represent a complete two-stage game of the type that we model above; the multiple repetitions allow subjects to gain experience with the game, potentially allowing convergence to equilibrium play. We suppose that people in the real world should be similarly experienced, having been involved with many partnerships and groups, some of which face "endgame play" issues.

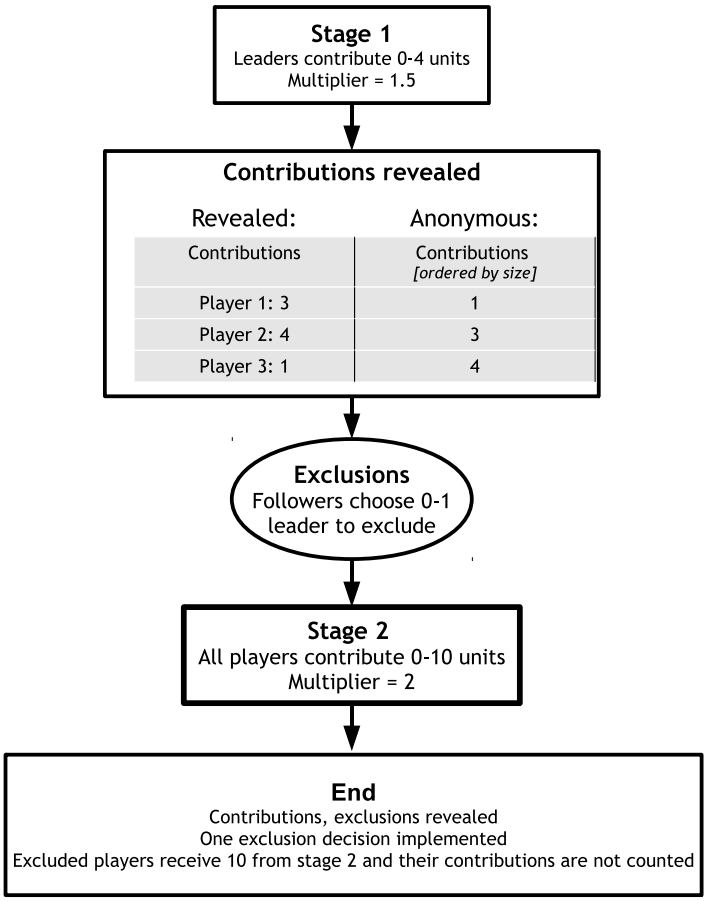
[14] If we had alternately *induced* a distribution of material conditionally cooperative payoffs among subjects, we would be merely testing their ability to signal strategically in a conventional setting. It would be difficult to argue that these material payoffs mimicked the relevant distribution of social preferences and subjects' beliefs over others types. Furthermore, strategic play over *social* preferences is not guaranteed to follow the same rules as for material preferences. Strategic play may itself change the nature of social and psychological rewards.

[15] Subjects were given separate sets of instructions for each treatment. Within each treatment two sets of instructions were given out, varying the amounts contributed in a worked example. These differing examples had significant effects (see online appendix 5). However, as these instruction treatments were administered orthogonally to our treatments and controlled for in our analyses, they do not bias our main results.

[16] The pilot version had six repetitions. The last of these repetitions involved choices specified using the strategy method. As we decided not to use the strategy method in subsequent runs, we use only the first five repetitions of the pilot in our data analysis below. None of our results change in sign if data from the pilot is excluded, and significance is preserved in most cases; tables available by request.

[17] Given our limited resources and the number of repetitions, we could not implement a "perfect strangers" design – two subjects may be in the same group in more than one repetition. However, as subject numbers change in each repetition, subjects can never know for certain who is in their group in a particular repetition. We address the possibility of session-specific and cross-repetition effects in section 4.4.

Figure 1: Experimental design



A repetition, depicted in Figure 1, consists of two linear VCMs. Players in each group are randomly numbered from 1 to 5; a player's number will vary from repetition to repetition. In the first game, players 1-3 (the "leaders")[18] play a smaller-stakes VCM game among themselves: each leader donates between 0 and 4 ECU's, and the total is multiplied by 1.5 and shared equally among the leaders. All players observe the leaders' contributions. Players 4-5 (the "followers") each vote either to exclude a specific leader – identified by player number – from the second stage, or to exclude no one. One of the followers' exclusion decisions is randomly selected and is implemented.

The distinction between leaders and followers is not part of our theory: it was implemented so as to prevent direct reciprocity motives from affecting either the decision to exclude or the second-stage contributions (of followers). This modification preserves the intuition behind our theory, as the followers will have similar incentives as leaders to use exclusion to screen out self-interested types.

The only difference between treatments is as follows. In the Revealed treatment, followers observe

[18]The terms "leader" and "follower" were not used in the experiment itself. In even repetitions, players 1-2 were selected for each group, from a pool made up of the previous repetition's players 4-5; players 3-5 were selected for each group from the remaining players. This ensured a reasonable balance of leader/follower roles across subjects.

the amount each leader contributed along with her player number, and can exclude on this basis. In the Anonymous treatment, leader contributions are not linked to player numbers, and hence followers cannot target specific leaders for exclusion; i.e., the choice is essentially whether or not to exclude a randomly chosen leader.

In the second stage, all players play another, larger VCM game: each player donates between 0 and 10 ECU's and the total donated by non-excluded players is multiplied by 2 and shared among all non-excluded players. Excluded players make a contribution decision but this is ignored in calculating payoffs; instead excluded players simply receive their second-stage endowment of 10 ECU's. This implies that, with standard preferences, being excluded is never strictly preferable, and if others contribute a positive amount in stage 2, being excluded is always strictly worse, since one can always contribute nothing in the second stage and profit at least 10 ECU's. Empirically, over 88% of non-excluded players earned more than 10 ECU's in the second stage.

Finally, all players learn choices, profits, and exclusion decisions. Announcing exclusion decisions only at the *end* of a repetition allows us an additional data point per round, and ensures that second-round contribution decisions are made in a relatively homogeneous environment, as subjects can not yet be certain whether an exclusion has been made. Announcing the decisions earlier would have allowed us to control for subjects' expectations about exclusion, but might have led to extraneous effects such as resentment at another's exclusion or non-exclusion.

The exclusion decision may be based on both material and psychological motives. Because the public good is divided amongst the included players, voting for exclusion will be in a player's material self-interest under certain beliefs. Specifically, a follower will increase her expected income by voting to exclude a leader whenever she expects that leader to contribute less than $\frac{3}{4}$ of the average contribution of the other group members. In the Anonymous treatment, while exclusion of a random *player* could not affect expected payoffs, the followers are randomly choosing one *leader* to exclude, and first stage behavior may indicate that the expected contribution of a *leader* is significantly lower than the average.

Previous experiments have found that a greater marginal per capita return (MPCR) tends to increase contributions, as does a larger group size, holding the MPCR constant (hence increasing total returns). However, if the total return rate is kept constant as group size increases, the effect of the decrease in the MPCR dominates, and contributions decline (Ledyard, 1993). It is impossible to keep both MPCR and total return rate the same when a player is excluded, hence we compromise between the two concerns. We set our total return rates (of 1.5 and 2 in stage 1 and 2 respectively) to have the same MPCR for the first and second stage in the presence of exclusion ($\frac{1.5}{3} = \frac{2}{4} = 0.5$). If there is no exclusion the MPCR is slightly lower in stage 2 ($\frac{2}{5} = 0.4$).This difference is unlikely to drive our results. Firstly, the difference is small (0.4 versus 0.5), and subjects do not know whether an exclusion will take place when they are making their decision, so that the *expected* difference in MPCR between treatments is even smaller. Second, an exclusion is slightly more likely in the *Revealed* case. If subjects are aware of this then the expected second stage MPCR is higher in the Revealed treatment, which would presumably lead to *greater* contributions in the Revealed treatment – the opposite of what we find.

In repetitions 3, 7, 11, and 15 (and in repetition 5 of the pilot session) after the first stage, we elicited guesses about all other players' second stage contributions. Guesses are incentivized using a quadratic scoring rule. Where the guess is about a leader player, it may take into account the amount that this leader contributed in the first stage (in the Anonymous treatment, we elicited predictions for e.g., "the player who contributed 2 in the first stage"). We elicited guesses in only four repetitions so that we could control for any effect of the elicitation on second stage behavior (cf. Gächter and Renner, 2010). The effects of the "guessing stage" variable were small and insignificant (results available on request). At the end of the game, participants received their payoffs from two randomly chosen repetitions, one for each stage, and one participant was paid for one randomly chosen guess.[19] Screen shots of key stages are shown in the online appendix (programmed using ZTree, created by Urs Fischbacher, 2007).

We ran four sessions with fifteen repetitions and one shorter pilot session with five repetitions, on a total of 150 experienced subjects from the standard pool at the University of Jena, including 91 females and 59 males; most subjects were students, from a wide variety of disciplines. Demographics are shown in Table 10 in the appendix. The experiment lasted approximately one hour. Subjects were paid a show-up fee of 2.50 Euros in addition to the profits mentioned below. Payments were made privately at the end of the experiment.

4.3 Summary statistics and overview

Overview

Summary statistics are shown in Table 1. For both treatments, stage 1 contributions were within the range typically found in prior work (Ledyard, 1993). Exclusion was common in both treatments, but subjects in the Revealed treatment were significantly more likely to vote to exclude someone (Fisher's exact test: p=0.006). The final column shows the number of votes to exclude a particular leader; across treatments, a leader's overall exclusion probability was roughly 15%. In the Anonymous treatment, since the selection of whom to exclude is effectively random, we replace the actual number of votes against a player with one third of the total votes (0, 1, or 2) to exclude *any* player for the relevant repetition and group. This substitution, used in all tables below, reduces random noise, but our results are not sensitive to it.

[19] Subjects' earnings are depicted algebraically in the online appendix.

Table 1: Summary statistics

	St. 1 Contr.[1]	St. 2 Contr.	Voted to exclude[2]	Mean guess[3]	Votes to exclude (sim.)[1],[4]
Anonymous					
Mean	2.14	4.34	.377	4.32	.251
Std. Dev.	1.21	3.24	.485	2.37	.222
Median	2	5	0	5	.333
Obs.	585	975	390	270	585
Revealed					
Mean	2.42	2.89	.495	3.46	.330
Std. Dev.	1.02	2.7	.501	2.32	.610
Median	2	2	0	3.5	0
Obs.	585	975	390	270	585
Overall					
Mean	2.28	3.61	.436	3.89	.291
Std. Dev.	1.13	3.07	.496	2.38	.46
Median	2	3	0	4	0
Obs.	1170	1950	780	540	1170

Abbreviations: St. = Stage, Contr.=Contribution

[1] For leader subjects only.

[2] For follower subjects only.

[3] Mean prediction for leaders' stage 2 contribution (in stages where predictions were made).

[4] Votes against leader subject. Conditional expectation simulated for anonymous case; see text below.

4.4 Hypotheses

We derive a series of empirical hypotheses from our model and from previous experimental evidence.[20]

Hypothesis 1. *(a) The higher are stage 1 contributions, the less subjects will exclude. (b) Lowering one's contribution will increase the risk of being excluded more in the Revealed than in the Anonymous treatment.*

Justification*:* All other subjects benefit from excluding a subject who donates less than average.[21] If the stage 1 contributions are informative of stage 2 choices, the (expected) material incentive to exclude a subject will decrease in that subject's stage 1 contribution. [22]

Part (b) stems from the simple fact that under anonymity a leader cannot be targeted based on her contribution. Her decision therefore only affects the overall probability of an exclusion. If there is an

[20] As noted in section 4.1, as we rely on homegrown preferences and beliefs, we expect a range of subject types; thus these hypotheses are not all derived *directly* from our model, but by analogy.

[21] Note that we allow that there may be *some* exclusion in the Anonymous treatment. Thus first-stage play may reveal that the leaders' average type is worse than the population average (prior belief), implying that randomly excluding a leader could be expected to increase the share of good types included in the second stage.

[22] The same effect would be caused by a fairness or a justice motive, as subjects would prefer to punish subjects who were previously uncooperative, or who they expect will be uncooperative. However, we can rule a *direct* reciprocity motive, as we can isolate the exclusion decisions of followers, who were not included in stage 1.

exclusion, she will only be "hit" with 1/3 probability.

Hypothesis 2. *Given the sizes of each round, the first stage is not large enough to substantially separate types when contributions are revealed; it will separate types to a greater extent when contributions are anonymous. Thus,*

(a) the correlation between an individual's stage 1 and stage 2 contributions will be non-negative and greater in the Anonymous treatment than in the Revealed treatment,

(b) subjects' expectations of stage 2 contributions will respond more to stage 1 contributions in the Anonymous treatment than in the Revealed treatment, and

(c) subjects' predictions of leaders' stage 2 contributions will be more informative (i.e., explain a greater share of the variation in actual stage 2 contributions) in the Anonymous treatment than in the Revealed treatment.

***Justification**:* As Proposition 2 suggests, separating behavior is more likely when stage 1 is anonymous. With less than complete pooling in the Anonymous case, the stage 1 contribution will be informative about stage 2 contribution. There will be more in the Revealed treatment as a subject's contribution has a greater impact on her risk of being excluded (see Hypothesis 1). Part (b) will result if subjects anticipate Hypothesis 1, and their expectations reflect this. For part (c), as explained above, stage 1 contributions are likely to have more information content (about true preferences and thus likely stage 2 behavior) in the Anonymous case then in the Revealed case, and subjects will realize this.

Hypothesis 3. *(a) Subjects' stage 2 contributions will increase with their expectations of others' stage 2 contributions. (b) Under anonymity, subjects' stage 2 contributions will increase in others' stage 1 contributions.*

***Justification**:* Part (a) reflects conditionally cooperative preferences. Part (b) incorporates hypothesis 2 – that there will be some separation in the Anonymous case – and thus behavior in both stage 1 and stage 2 will reflect a subject's social preferences and beliefs.

Hypothesis 4. *The extra information in the Anonymous treatment will result in higher stage 2 contributions on average.*

***Justification**:* Since VCM experiments have consistently found that contributions decline over repetitions,[23] we suspect that the equilibrium without a credible institution to signal the presence of conditional cooperators will have low cooperation levels. This, we suspect that the ability to credibly signal this, in the Anonymous treatment, will increase stage 2 contributions relative to the less credible Revealed stage 1 contributions. However, this is our most speculative hypothesis, as our model predicts that "information helps" only under specific conditions.

[23] In explaining this pattern, Ostrom (2000) argues that conditional cooperators (good types) begin optimistic, and some selfish types strategically pool with them (as in Kreps et al., 2001). (As Holt and Laury (2008) note, the former group must "systematically overestimate" their prevalence). As the end of the game approaches and free riding occurs, the good types become disappointed, reducing their contributions and discouraging other good types from contributing. "Without ... institutional mechanisms to stop the downward cascade, eventually only the most determined conditional cooperators continue to make positive contributions in the final rounds." We see our anonymous first stage as one such mechanism.

4.5 Results

We test the above hypotheses on our experimental data. Result numbers correspond to hypothesis numbers. For subject-level results we examine individual behavior in a linear regression framework.[24] While observations at the treatment/session level are strictly independent, they do not fully exploit the information in the data. As in all experiments with imperfect stranger matching, the per-subject observations are not completely independent, and play may be affected by experience in earlier repetitions. To address this, we estimate robust standard errors, clustered at either the subject or the treatment/session level as appropriate. Where relevant we also control for subjects' experience in previous repetitions, or include session, treatment, or a subject-fixed effects.

Result 1. *Higher stage 1 contributions by a subject reduce the probability of the subject's exclusion in both treatments, but this effect is stronger in the Revealed treatment.*

[24] Earlier drafts used a Poisson exponential regression; we present the simpler specification here for greater transparency and comparability and more straightforward testing of interaction terms. We focus on our "key results" for a further set of robustness checks, available by request, namely: Table 3, column 2, Table 4, column 2, Table 6, columns 2, 3, and 5, and Table 9, column 1. These results are also preserved in sign under several alternate specifications (available by request), including Poisson, Tobit, and negative binomial (only Tobit is checked for the IV regression, and it is not used for the regression with fixed effects). For Table 4 column 2 and for Table 7 column 3, the results discussed are no longer significant in some specifications. Otherwise, significance is preserved at at least $p<0.10$ in all cases, both for raw coefficients and when we estimate marginal effects over all values of the independent variables. Details and regression tables are available by request.

Table 2: Linear regressions: exclusion votes against a subject

Dependent variable = Number of votes to exclude subject (in single repetition).[1]			
	(1)	(2)	(3)
	All Repetitions	Repetitions 8-15	All Repetitions
Stage 1 Contribution	-0.033*	-0.045*	
	(0.010)	(0.014)	
Revealed × Contibution	-0.34**	-0.35**	
	(0.024)	(0.034)	
Repetition	0.0038	-0.00072	0.0072+
	(0.0033)	(0.0064)	(0.0033)
Revealed × Repetition	-0.0036	0.014	-0.0039
	(0.0038)	(0.014)	(0.0056)
Contributed 1 ecu			0.093
			(0.11)
Revealed × Contributed 1 ecu			-0.61**
			(0.16)
Contributed 2 ecu's			0.084
			(0.10)
Revealed × Contributed 2 ecu's			-1.29**
			(0.15)
Contributed 3 ecu's			0.041
			(0.11)
Revealed × Contributed 3 ecu's			-1.51**
			(0.16)
Contributed 4 ecu's			0.039
			(0.10)
Revealed × Contributed 4 ecu's			-1.60**
			(0.16)
Session/Treatment Dummies	Yes	Yes	Yes
Revealed Dummies	1.2**	0.9**	1.5**
Observations	1170	576	1170

[1] Conditional expectation simulated for anonymous case; see top of section 4.3.
Robust (clustered by session/treatment) standard errors in parentheses.
'Revealed Dummies.' gives average of intercepts for revealed session-treatments
+ p<0.10, * p<0.05, ** p<0.01

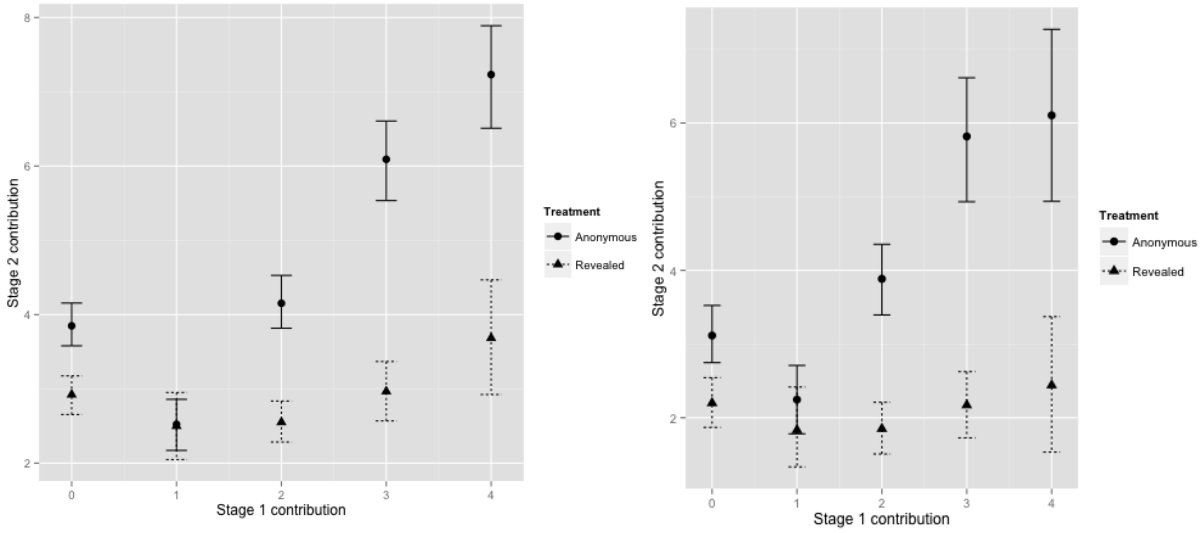
Table 2 demonstrates that the probability a leader was excluded varied inversely with her stage 1 contribution and this effect was much stronger in the Revealed treatment. [25] The effect appears to be nonlinear and present both the intensive and extensive margins. Overall in the Revealed treatment, given stage 1 contributions of 0,1,2,3, and 4, the probabilities of exclusion were 72%, 48%, 13%, 7%, and 5% respectively.

Result 2. *(a) A leader's stage 1 contribution is a better predictor of her stage 2 contribution in the Anonymous treatment then in the Revealed treatment.*

[25] We cluster errors at the session-treatment level; as a subject can not be identified by other subjects by her behavior in earlier repetitions, there should be no subject-specific error term here.

Figure 2: Leader's mean stage 2 contribution by her stage 1 contribution.

(a) Overall (b) Repetitions 8-15



Note: the bars show bootstrapped 95% confidence intervals for the conditional mean.

Figure 2 shows a leader's mean stage 2 contribution by her stage 1 contribution for each treatment. We present these both overall and for later repetitions, presumably after some strategic learning. Both graphs show a positive correlation between giving in the stages, which appears much stronger in the Anonymous treatment. Next we decompose the variance into its explained and unexplained components, reporting marginal and total sums of squares.[26]

[26]Interpreting the ANOVA in a regression framework, the marginal sum of squares can be interpreted as "the reduction in R-sq if you removed that variable only." These add up to the total sum of squares (TSS) only if the variables are exactly orthogonal. Regression analysis of stage 2 contributions obviously yielded comparable results, available by request.

Table 3: Analysis of variance of stage 2 contributions by stage 1 contributions

Partial (marginal) sums of squares:								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Anon.	Revealed	Anon. later[1]	Rvld. later[1]	Anon.	Rvld.	Anon, not min[2]	Rvld, not min[2]
1 ecu Contr.	14	14	35**	7	16	21*	254**	7.4**
2 ecu Contr.	211**	20	199**	9.1	37+	24*	558**	126**
3 ecu Contr.	605**	42+	428**	15+	36*	28*	131*	41
4 ecu Contr.	1008**	88*	497**	19+	75**	44**	.	.
Subj. effects					2038**	1560**		
Model DF	4	4	4	4	78	78	66	3
Observations	585	585	288	288	585	585	263	247
Model SS	1937	133	849	27	3975	1694	1743	132
Total SS	6172	3997	2846	1519	6172	3997	2911	2268
R-sq.	.31	.033	.3	.018	.64	.42	.6	.058

Abbreviations: Anon. = Anonymous, Rvld= Revealed, Contr.=Contribution, St.=Stage, Subj.=Subject

[1] 'Later' refers to repetitions 8-15.

[2] 'Not min' removes subjects who contributed the lowest amount in that first stage.

+ p<0.10, * p<0.05, ** p<0.01; tests from analogous regressions clustered on id

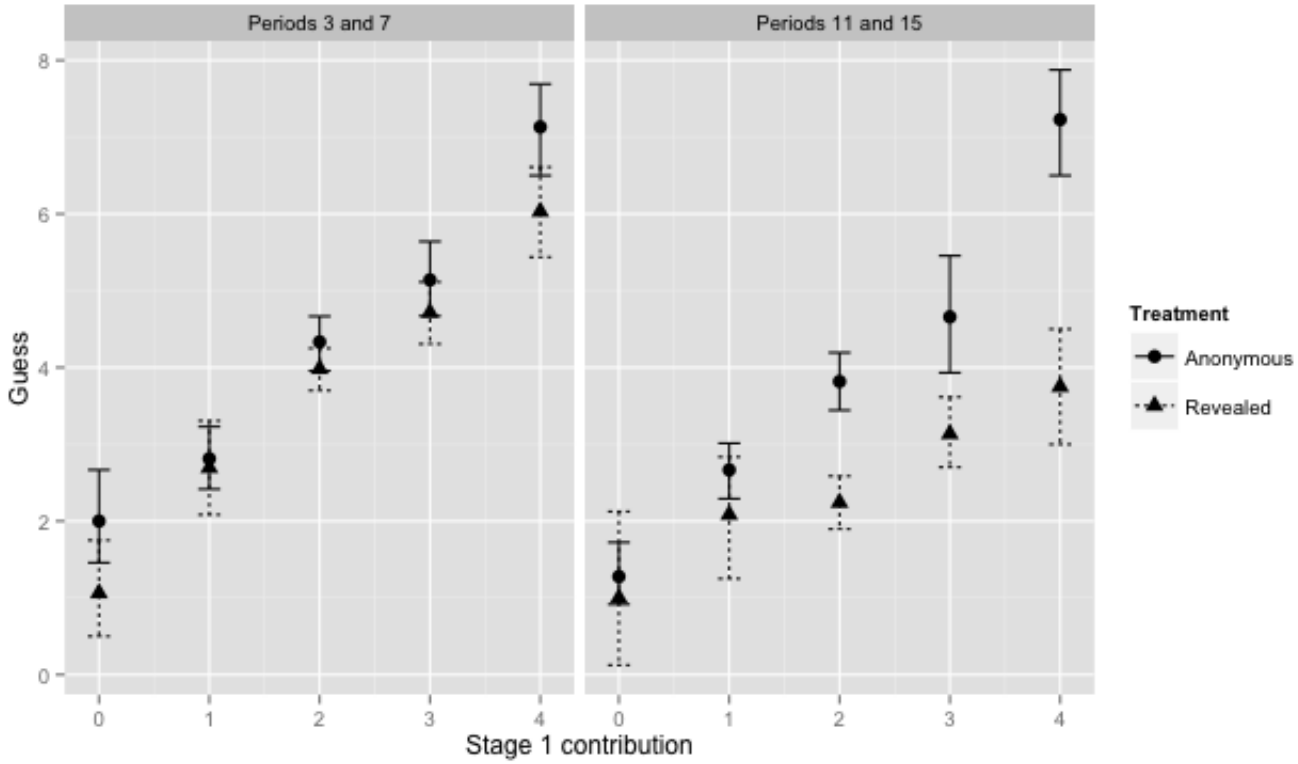
A subjects' first stage contribution explains much of the variance in her stage 2 contribution in the Anonymous treatment, while in the Revealed treatment it explains very little. It is not just the *presence* of an anonymous stage 1 contribution that matters, but also its magnitude; a 1 ECU contribution explains little, while larger contributions matter a great deal.

Almost all the explanatory power of first stage contribution is via individual heterogeneity. In columns 5 and 6, after conditioning on subject-specific effects, first stage contribution explains little of the remaining variation for either treatment. That is, first stage contributions "explain" second stage contributions in the Anonymous treatment because they are reliable signals of the individual leaders' types. As *subjects* cannot observe the identity or type of the leaders in their group, this signal is important for them.

An alternative explanation is that subjects who contributed the lowest amount in stage 1 might expect to be excluded, and thus might not take their stage 2 choice as seriously. The final two columns of Table 3 remove the leaders who contributed least in a repetition. Even with these removals, which are approximately balanced across treatments, stage 1 contribution still has more explanatory power in the Anonymous treatment than in the Revealed treatment.

Result 2. (b) and (c) *Others' predictions about a subject's stage 2 contributions are positively correlated to that subject's stage 1 contributions. This effect is stronger in the Anonymous treatment than in the Revealed treatment, and these predictions are more accurate.*

Figure 3: Leaders' mean predicted stage 2 contributions by their actual stage 1 contributions



Note: the bars show bootstrapped 95% confidence intervals for the mean. "Guess" is mean prediction for leader's stage 2 contribution.

Figure 3 shows subjects' mean predictions of leaders' stage two contributions, plotted against the leaders' actual *first* stage contributions. subjects in the Revealed treatment become more skeptical in later stages, and predicted contributions become much lower for high stage 1 contributors. Table 4 formalizes this, regressing subjects' predictions about leaders' stage 2 contributions on the leaders' actual stage 1 contributions.[27] The coefficient of first-stage contribution is significantly lower in the Revealed case, although the summed coefficient remains significant.

[27] In the Anonymous treatment predictions were for (e.g.) "the subject who contributed 4 ECU's."

Table 4: Subjects' predictions for leaders

Dependent variable = Subject's prediction of target's stage 2 contribution				
	(1)	(2)	(3)	(4)
	All repetitions	...	Repetitions 11,15	...
Target Stage 1 Contribution.	1.39**	1.08**	1.40**	1.03**
	(13.16)	(9.49)	(9.93)	(6.34)
Target Stage 1 Contribution × Revealed	-0.44**	-0.30+	-0.69**	-0.44+
	(-2.81)	(-1.96)	(-2.99)	(-1.84)
Repetition	-0.029	0.022		
	(-1.12)	(0.83)		
Revealed × Repetition	-0.18**	-0.15**		
	(-4.74)	(-4.14)		
Revealed Treatment	1.53**	1.59**	-0.20	0.60
	(3.22)	(3.28)	(-0.39)	(0.99)
Constant	1.50**	-0.78+	1.14**	-0.83*
	(4.68)	(-1.74)	(4.49)	(-2.05)
History & Lag 1 Variables	No	Yes	No	Yes
Observations	1152	1152	576	576
Sum coefficient: Contribution & Revealed	0.95**	0.78**	.71**	.59**

(Pilot) Session 1 excluded because of fewer prediction rounds.

Robust standard error (clustered by subject) in parentheses.

In anonymous treatments predictions were for (e.g.,) <the person who contributed 4 ecus>.

+ p<0.10, * p<0.05, ** p<0.01

As anticipated, subjects' predictions for others' contributions were more accurate in the Anonymous than in the Revealed treatment, and this difference was significant for predictions made in the later repetitions, as shown in Table 5.[28]

Table 5: Subjects' prediction accuracy and bias

Treatment	*Anon.*	*Revealed*	*Difference*$^{(B)}$	*Anon*	*Revealed*	*Dfc.*$^{(B)}$
Repetitions	*3,5,7,11,15*	*3,5,7,11,15*	*3,5,7,11,15*	*11,15*	*11,15*	*11,15*
Mean Absolute error$^{(A)}$	2.43**	2.43**	0.00	2.08**	2.44**	-0.36*,*
(std. error)	(.08)	(.08)	(.12)	(.12)	(.13)	(.17)
Corr(predicted, actual)	0.39**	0.26**	[.19]$^{(C)}$	0.52**	0.10	[.00**]$^{(C)}$

+ p<0.10, * p<0.05, ** p<0.01, 2-tailed significance tests (t-tests, robust to allow unpaired data to have unequal variance).

Marks after comma: Significance in 2-tailed rank-sum tests.

(A) "Error" is subject's prediction for another subject less that subject's actual contribution. Results for mean square errors (by request) are similar.

(B) "Difference" is for previous two columns, i.e., value for anonymous minus value for Revealed treatment.

(C) Brackets: P-value for difference in correlations (significance from corresponding linear regression, robust standard errors clustered by id).

[28]Because of this poor predictability, in the Revealed treatment, the "number of votes to exclude a subject" was a poor predictor of this subject's stage 2 contribution choice. As a result, these targeted exclusions did not directly increase stage 2 payoffs for the remaining subjects. These results are given in the Online Appendix.

Result 3. *(a) A follower's stage 2 contribution increases in her expectation of others' stage 2 contribution. (b) In the Anonymous treatment there is a significant positive relationship between a follower subjects' stage 2 contribution and the average stage 1 contribution she observed in that repetition. This holds in both earlier and later repetitions.*

We investigate this in Table 6. In each of these regressions, we only include followers to rule out motives such as direct reciprocity for stage 1 or bitterness from the perceived probability of being excluded.[29]

Table 6: Determinants of followers' stage 2 contribution

	(1) All Repetitions	(2) Repetitions 8-15	(3) FE: All Repetitions
Average (others') Stage 1 Contribution	0.83**	0.83*	-0.65
	(0.30)	(0.40)	(0.80)
Revealed × Average (others') Stage 1 Contribution	0.030	-0.587	0.365
	(0.426)	(0.535)	(0.914)
Dummy: Revealed Treatment	-0.194	2.829	
	(1.070)	(1.963)	
Repetition	-0.158**	-0.138	-0.001
	(0.040)	(0.099)	(0.145)
Revealed × Repetition	-0.074	-0.206	0.147
	(0.052)	(0.126)	(0.166)
Average prediction for leaders			0.784+
			(0.426)
Revealed × Average prediction			0.626
			(0.491)
Constant	2.951**	2.946+	-0.257
	(0.781)	(1.575)	(1.731)
Additional controls	Yes	Yes	No
Observations	780	384	192
Sum coefficient: Stage 1 Contribution, Revealed	0.856**	0.245	-0.287

Robust (clustered by subject) standard errors in parentheses.
Additional controls: High example dummy, high example-revealed interaction, last repetition, revealed × last repetition.
In anonymous treatments predictions were for (e.g.,) <the person who contributed 4 ecus>.
+ p<0.10, * p<0.05, ** p<0.01

The first two columns of Table 6 measure the relationship between a follower's second stage contribution and leaders' average first stage contributions, presumably arising through the conditional cooperation motive. Column 1 reveals a significant positive relationship between a follower subjects' stage 2 contribution and the average stage 1 contribution she observed in that repetition. This (result 3b) supports hypothesis 3b. While the positive relationship holds for both treatments, in column 2 we see that in later repetitions this effect is substantially smaller and insignificant in sum for the Revealed

[29] Still, even if we include leaders, in columns 1 and 2 the coefficients on "average contribution" in the Anonymous treatment remains positive and significant, and the adjustment to this coefficient in the Revealed treatment becomes *more* negative and significant in column 2. In the remaining columns the coefficients on "average guess" remain positive and significant. Details are available by request.

treatment, presumably reflecting the lower informativeness.

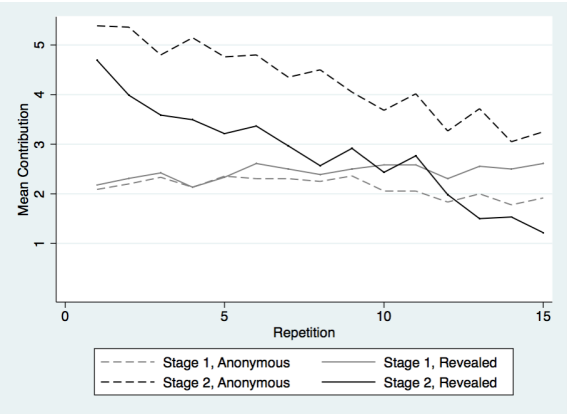
In column 3 we include the average of a subject's prediction for leaders' contributions (interacted with the treatment) as a regressor. Because this prediction may be correlated with subject-specific unobservables (e.g., more generous people may be more optimistic about others) we control for a subject-specific effect. After controlling for beliefs, others' first-stage contributions no longer have a positive significant effect. This suggests that the effect works through *inferences* about second-stage contributions (rather than as a *direct* response to the leaders' contributions). There is clear evidence of conditional cooperation: the coefficients on the average prediction are positive and significant.[30] This causal "conditional cooperation" interpretation might still be criticized, for example, because the subject may first choose how much to contribute and her prediction may be an ex-post rationalization of this choice (Fehr and Schmidt 2006). In response, we first note that our subjects' guesses are financially motivated. Secondly, papers such as Chaudhuri and Paichayontvijit (2011) and Smith (2012) support the "conditional cooperator" interpretation. Finally, as an additional test of this interpretation, we ran instrumental variables regressions (see table 8, presented and discussed in the Appendix "Robustness Checks") which strongly support this intepretation.

Result 4. *Stage 2 contributions, and overall earnings, are significantly higher in the Anonymous treatment.*

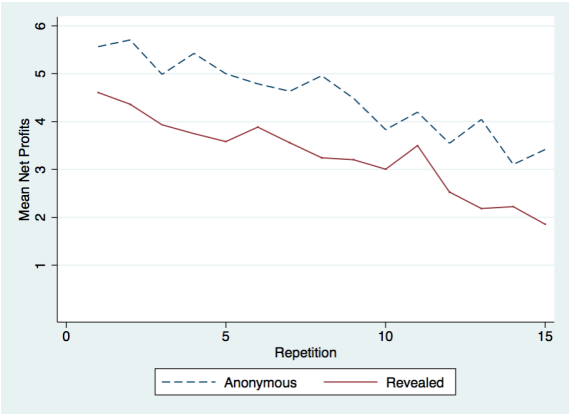
This is demonstrated in the figures and tables below.

Figure 4: Mean contributions, net profits by repetition by treatment

(a) Contributions



(b) Net Profits

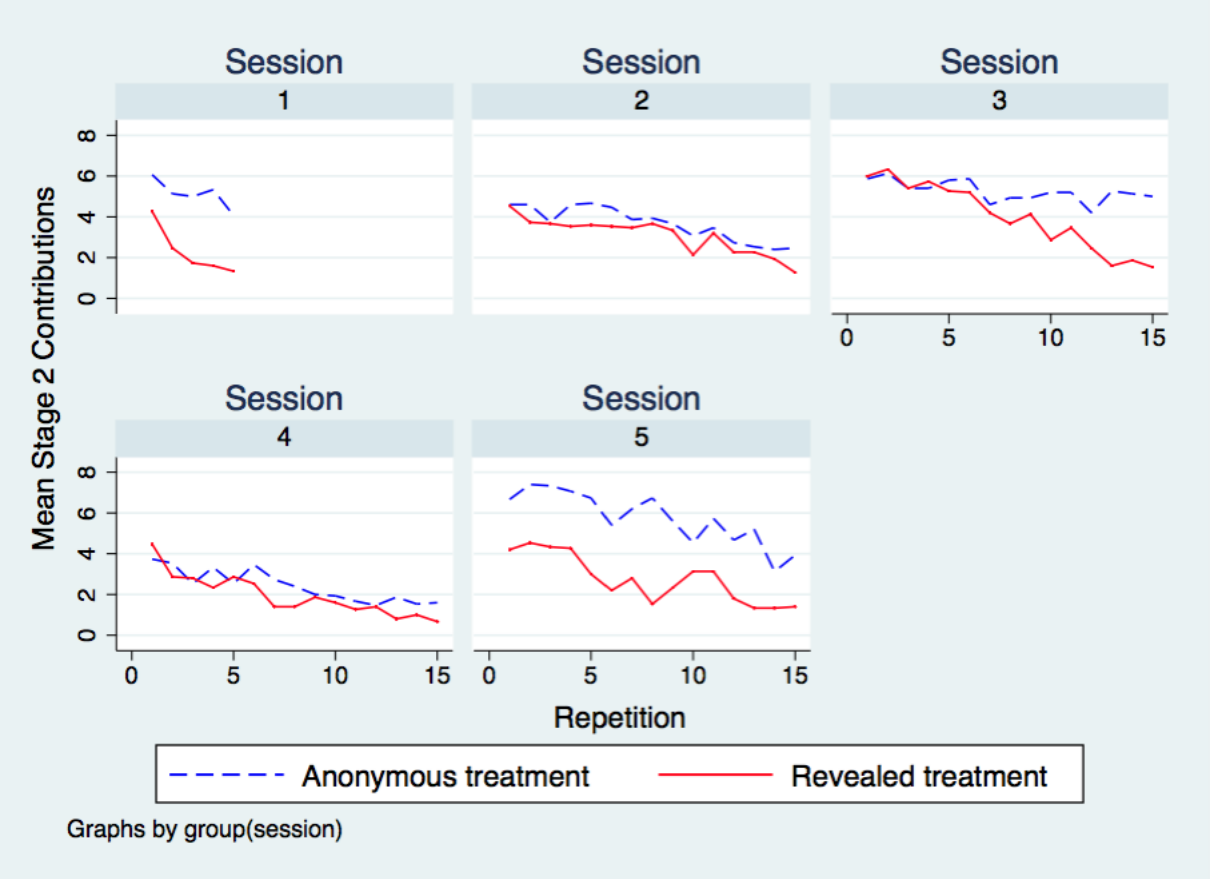


Note: "Net Profits" excludes rewards from accurate predictions.

[30]The coefficient is larger in the Revealed treatment; this does not contradict our hypothesis, which implied that the slope of *beliefs in stage 1 contributions* would be smaller, but did not refer to the slope of *stage 2 contributions in beliefs*. We do not include lagged controls for a subject's experience in previous repetitions here, as we expect the effect of these to be subsumed in the subjects' expectations; the results (by request) are not sensitive to this. For all columns, regressions including all four of a subject's predictions for other subjects yielded similar results (by request). "Endgame" effects were not significant, and mostly (insignificantly) more negative for the Revealed treatment.

As Figure 4 shows, average stage 1 contributions began very similar for each treatment, and remained fairly constant across repetitions, ending somewhat lower in the Anonymous treatment. In contrast, in the Anonymous treatment stage 2 average contributions began over 50% and remained above 30%, while in the Revealed treatment they began slightly below 50% and declined to less than 15%. Figure 5 shows that these patterns are fairly similar across sessions, although there was some variation.

Figure 5: Mean stage 2 contributions by session and treatment



Note: Only five comparable repetitions were run in the pilot session (session 1).

In Table 7 we see that the difference between treatments in average stage 2 contributions is statistically significant in rank sum tests at the treatment/session level.[31]

Additional result: *Stage 1 contributions are significantly higher in the Revealed treatment.*

Stage 1 contributions are higher in the Revealed treatment; significance tests are given in Table 7; this difference is significant at the subject level only, and particularly in later repetitions. Our theory predicts that for an intermediate size first stage, bad types will pool with good types in stage 1 of the

[31] These results are approximately equivalent in t-tests, available by request. The effect is especially strong in later repetitions, where there has been the most learning, and in the final repetition, where there is no possibility of influencing potential future partners' play in random stranger matchings. The latter differences are significant in t-tests but not in all of the above rank-sum tests.

Revealed treatment but they will separate in the Anonymous treatment. If selfish types learn to pool in the Revealed treatment, then first stage contributions will end up higher in that treatment; only pooling on "high" (rather than "low") contributions will survive the intuitive criterion.

In net, weak enforcement was materially beneficial to subjects: average net profits were 4.58 Euros in the Anonymous treatment and 3.35 Euros in the Revealed treatment; this difference is strongly significant at the subject level (see Table 7). These results show that, at least for one environment, anonymity and weak enforcement help to increase efficiency. The previously presented evidence is consistent with our model, suggesting that this was driven by conditional cooperation and more accurate signals of others' likely second-stage contributions.

Table 7: Rank sum tests – effect of Anonymous treatment on contributions and earnings

St. 2 mean contr., Prob(anon>revealed)	Session-Treatment level	Subject level
Overall	0.840+	0.697**
	(0.076)	(0.000)
No pilot session	0.750	0.654**
	(0.248)	(0.004)
Later repetitions (>7)	0.812	0.669**
	(0.149)	(0.001)
Last repetition, no pilot session	1.000*	0.430
	(0.021)	(0.166)
St. 1 mean contr., Prob(anon>revealed)	Session-Treatment level	Subject level
Overall	0.380	0.412+
	(0.530)	(0.062)
No pilot session	0.312	0.375*
	(0.387)	(0.019)
Later repetitions (>7)	0.312	0.358**
	(0.387)	(0.007)
Overall earnings, Prob(anon>revealed)	Session-Treatment level	Subject level
Overall	0.640	0.533*
	(0.465)	(0.012)
No pilot session	0.563	0.522
	(0.773)	(0.107)
Later repetitions (>7)	0.563	0.515
	(0.773)	(0.416)

Rank sum tests for average contribution by session/treatment (left side), and a subject's average contribution (right side).

Values: probability a randomly chosen Anonymous average contribution ranks above a randomly chosen Revealed average contribution.

P values in parentheses.

+ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$

Abbreviations: St. = stage, contr = contribution, anon = anonymous

While the evidence presented suggests that our theory explains the greater contributions in the Revealed treatment, it is also possible that in this treatment, excluded subjects became more resentful, and thus contributed less in later repetitions, sparking the rapid decline. In Appendix A.2 we test for

this alternative in several ways, and find no evidence that greater embitterment in the Revealed case is driving our results.

5 Conclusion

The experimental results are consistent with our theory. The anonymous first round appears to have helped players learn about the preferences of their fellow group members. In the Anonymous treatment leaders contributed what they really wanted to contribute, while in the Revealed treatment they largely contributed to avoid exclusion. The lower uncertainty about types in the Anonymous treatment lead to greater efficiency: second stage contributions were (marginally) significantly and substantially higher, as were overall profits.[32] The robust positive relationship between subjects' contributions and their predictions of others' contributions supports our "conditional cooperation " explanation for this.

As our model shows, greater certainty need not lead to greater contributions, even where subjects are conditional cooperators. This will depend on whether the benefit when "good news" is revealed outweighs the cost when "bad news" is revealed. This appears to hold in our experiment. We suspect that many subjects are loath to contribute when they predict even a small chance that others do not. In our "low information" Revealed treatment, in the absence of credible signals, this small risk may be ever-present, and may be stifling contributions.[33]

Until now, punishment and exclusion have been seen as increasing public goods provision. We show that they can do the reverse. Cooperative behavior has been found to emerge in the context of local "low enforcement" institutions and moral norms (Ostrom, 2000; Cardenas et al., 2000). When it does, this may be informative about the pro-social preferences or "types" of the participants. Stronger enforcement and punishment institutions can destroy this information, by forcing everyone to behave well. This may destroy trust in others' true willingness to contribute, which may lead to less cooperation when the institutions cease to be available. The gains from making enforcement *weaker* are likely to be large when "final stage" cooperation, which cannot be enforced by the threat of subsequent sanctions, is important. For instance, episodes of conflict, natural disasters, and economic crises all put a premium on groups' ability to cooperate, and simultaneously make the environment uncertain so that future group interactions cannot be guaranteed.

Finally, we note that our results have implications for policy. Karlan (2005) finds that trustworthiness in a laboratory trust game (run on Peruvian microcredit participants) predicts repayment of a loan "enforced almost entirely through social pressure." Games like this, with anonymous participation, might help participants build trust (in cases where it is warranted) and lead to greater contributions to lending pools and other collective goods. Future field experiments should explore this possibility.

[32] We admit that significance is marginal for our session-level analysis for hypothesis 4. Our evidence for hypotheses 1-3 – which are also more strongly grounded in theory – is stronger.

[33] However, as we did not elicit the subjects' entire distribution of beliefs (only the predicted means), we can not isolate the channel through which this occurs. In the online appendix we further discuss explanations in the context of previous experimental work.

References

Ahn, T., R. Isaac, and T. Salmon (2009). Coming and going: Experiments on endogenous group sizes for excludable public goods. *Journal of Public Economics 93*(1-2), 336–351.

Alpizar, F., F. Carlsson, and O. Johansson-Stenman (2008). Anonymity, reciprocity, and conformity: Evidence from voluntary contributions to a national park in Costa Rica. *Journal of Public Economics 92*(5-6), 1047–1060.

Andreoni, J. and R. Petrie (2004). Public goods experiments without confidentiality: a glimpse into fund-raising. *Journal of Public Economics 88*, 1605–1623.

Bochet, O., T. Page, and L. Putterman (2006). Communication and punishment in voluntary contribution experiments. *Journal of Economic Behavior & Organization 60*(1), 11–26.

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

Cardenas, J., J. Stranlund, and C. Willis (2000). Local environmental control and institutional crowding-out. *World Development 28*(10), 1719–1733.

Carpenter, J. and C. K. Myers (2010). Why volunteer? evidence on the role of altruism, image, and incentives. *Journal of Public Economics 94*(11), 911–920.

Chaudhuri, A. (2011). Sustaining cooperation in laboratory public goods experiments: a selective survey of the literature. *Experimental Economics 14*(1), 47–83.

Chaudhuri, A. and T. Paichayontvijit (2011). Does strategic play explain the decay in contributions in a public goods game? experimental evidence. *Mimeo*.

Cinyabuguma, M., T. Page, and L. Putterman (2005). Cooperation under the threat of expulsion in a public goods experiment. *Journal of Public Economics 89*(8), 1421–1435.

Cooter, R. and B. Broughman (2005). Charity, Publicity, and the Donation Registry. *The Economists' Voice 2*(3), 4.

Coricelli, G., D. Fehr, and G. Fellner (2004, June). Partner selection in public goods experiments. *Journal of Conflict Resolution 48*(3), 356–378.

De Oliveira, A., R. Croson, and C. Eckel (2009). One bad apple: Uncertainty and heterogeneity in public good provision.

Dixon, M. and V. J. Roscigno (2003). Status, networks, and social movement participation: The case of striking workers1. *American Journal of Sociology 108*(6), 1292–1327.

Ehrhart, K. M. and C. Keser (1999). *Mobility and Cooperation: On the Run*. Centre interuniversitaire de recherche en analyse des organisations.

Ermisch, J. and D. Gambetta (2010). Do strong family ties inhibit trust? *Journal of Economic Behavior & Organization 75*(3), 365–376.

Fehr, E. and S. Gächter (2000). Cooperation and punishment in public goods experiments. *The American Economic Review 90*(4), 980–994.

Fehr, E. and K. M. Schmidt (2006). The economics of fairness, reciprocity and altruism–experimental evidence and new theories. *Handbook on the economics of giving, reciprocity and altruism 1*, 615–691.

Fischbacher, U. (2007). z-Tree: Zurich toolbox for ready-made economic experiments. *Experimental Economics 10*(2), 171–178.

Fischbacher, U., S. Gächter, and E. Fehr (2001). Are people conditionally cooperative? Evidence from a public goods experiment. *Economics Letters 71*(3), 397–404.

Friedman, M. (1953). The methodology of positive economics. *The Philosophy of economics: an anthology 2*, 180–213.

Fudenberg, D. and E. Maskin (1986). The folk theorem in repeated games with discounting or with incomplete information. *Econometrica: Journal of the Econometric Society*, 533–554.

Gächter, S. and E. Renner (2010). The effects of (incentivized) belief elicitation in public goods experiments. *Experimental Economics 13*(3), 364–377.

Glazer, A. and K. Konrad (1996). A signaling explanation for charity. *American Economic Review 86(4)*, 1019–1028.

Harbaugh, W. T. (1998, May). The prestige motive for making charitable transfers. *The American Economic Review 88*(2), 277–282.

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

Holt, C. and S. Laury (2008). Theoretical explanations of treatment effects in voluntary contributions experiments. C. Plott, V. Smith, eds. *Handbook of Experimental Economic Results*.

Hugh-Jones, D. and D. Reinstein (2012). Anonymous rituals. *Journal of Economic Behavior and Organization 81*(2), 478.

Ichino, A. and G. Muehlheusser (2008). How often should you open the door?: Optimal monitoring to screen heterogeneous agents. *Journal of Economic Behavior & Organization 67*(3), 820–831.

Isaac, R. M., J. Walker, and S. Thomas (1984). Divergent evidence on free riding: An experimental examination of possible explanations. *Public Choice 43(1)*, 113–149.

Karlan, D. (2005). Using experimental economics to measure social capital and predict financial decisions. *American Economic Review 95*(5), 1688–1699.

Katz, E. and J. Rosenberg (2005). An economic interpretation of institutional volunteering. *European Journal of Political Economy 21*(2), 429–443.

Katzenbach, J. R., D. K. Smith, and D. Smith (2001, April). *The Discipline of Teams: A Mindbook-Workbook for Delivering Small Group Performance* (1 ed.). Wiley.

Keser, C. and F. van Winden (2000). Conditional Cooperation and Voluntary Contributions to Public Goods. *Scandinavian Journal of Economics 102*(1), 23–39.

Kreps, D., P. Milgrom, J. Roberts, and R. Wilson (2001). Rational Cooperation in the Finitely Repeated Prisoners Dilemma. *Readings in Games and Information*.

Ledyard, J. (1993). *Public Goods: A Survey of Experimental Research*. Division of the Humanities and Social Sciences, California Institute of Technology.

Londregan, J. and A. Vindigni (2006). Voting as a credible threat.

Martin, R., P. Fosh, H. Morris, P. Smith, and R. Undy (1991). The decollectivisation of trade unions? ballots and collective bargaining in the 1980s. *Industrial Relations Journal 22*(3), 197–208.

Milinski, M., D. Semmann, and H. Krambeck (2002). Reputation helps solve the'tragedy of the commons'. *Nature 415*(6870), 424–6.

Ostrom, E. (2000). Collective action and the evolution of social norms. *The Journal of Economic Perspectives*, 137–158.

Page, T., L. Putterman, and B. Unel (2005). Voluntary association in public goods experiments: Reciprocity, mimicry and efficiency. *Economic Journal 115*(506), 1032–1053.

Plott, C. and V. Smith (2008). *Handbook of results in experimental economics*. North-Holland.

Rotemberg, J., G. Saloner, and R. Oligopoly (1986). A supergame-theoretic model of price wars during booms. *New Keynesian Economics 2*, 387–415.

Sass, M. and J. Weimann (2012). The dynamics of individual preferences in repeated public good experiments. Technical report, Otto-von-Guericke University Magdeburg, Faculty of Economics and Management.

Sitzia, S. and R. Sugden (2011). Implementing theoretical models in the laboratory, and what this can and cannot achieve. *Journal of Economic Methodology 18*(4), 323–343.

Smith, A. (2012). Estimating the causal effect of beliefs on contributions in repeated public good games. *Experimental Economics*.

Soetevent, A. (2005). Anonymity in Giving in a Natural Context: An Economic Field Experiment in Thirty Churches. *Journal of Public Economics 89*(11-12), 2301–2323.

Spence, M. (1973). Job Market Signaling. *Quarterly Journal of Economics 87*(3), 355–374.

Veblen, T. (1899). *The theory of the leisure class: an economic study in the evolution of institutions*. Macmillan, New York.

A Appendix

A.1 Proofs

Notation

Define $\beta(x)$ as the best response to good-type donations of x when the distribution of good types is same as the prior. Hence, $\beta(x)$ satisfies

$$\sum \Pi(g) w_1(\beta(x), \frac{gx}{N-1}) = 1 - \bar{\alpha}. \tag{6}$$

Our conditions on w ensure that $\beta(0) > 0$ and $\beta'(x) > 0$ on $x \in [0, \bar{x}]$.

Proof that all equilibria are symmetric among good types:

Proof. When others' donations are uncertain, good types' optimal donations satisfy (4). Since w is strictly concave, this has a unique solution. Thus no good type plays a mixed strategy in equilibrium.

Suppose first that the number of good types is revealed to be $g+1$. We consider two cases: either the identity of the good types is known, or it is completely unknown. (These correspond to our revealed and anonymous signaling institutions.) Write π_{ij} for the probability, believed by any good type j, that player i is good. Players know their own type, so $\pi_{ii} = 1$. For other players, in the first case, $b\pi_{ij} = 1$ for $i \in G$, a set of g players, and 0 for all others; in the second case, $\pi_{ij} = \frac{g}{N-1}$ for all $i \neq j$. In either case, $\pi_{ij} = \pi_{ik}$ for $i \neq j$, $i \neq k$, $j \neq k$; pairs of good types share common probabilities about the type of third players. Also note that for $j \neq k$, $\pi_{jk} = \pi_{kj}$ if $j,k \in G$ in the identity known case, and always if identities are unknown.

Say that in equilibrium player i plays x_i if he is a good type. Let $j = \arg\max_i x_i$ and $k = \arg\min_i x_i$ (the maximum and minimum being taken over G in the identity known case). Suppose for a contradiction that $x_j > x_k$. Player j best responds to his expected distribution of others' donations, solving

$E_{\bar{X}_{-j}} w_1(x_j, \bar{X}_{-j}) = 1 - \bar{\alpha}$, where $\bar{X}_{-j}$ is derived from the probabilities π_{ij} and donations x_i for all $i \neq j$. Similarly player k best responds to $\bar{X}_{-k}$. Now, since $\pi_{jk} = \pi_{kj}$, $x_j > x_k$ and all other probabilities and donations are common to both i and j, the distribution $\bar{X}_{-j}$ is first order stochastically dominated by $\bar{X}_{-k}$. But then, by $w_{12} > 0$ and $w_{11} < 0$, it must be that $x_j < x_k$, a contradiction.

The proof when the number of good types is unknown is similar and is omitted. □

Proof that x^* exists and is unique, and x_g exists and is unique for any g:

Proof. Suppose first that the number of good types is known to be $g+1$. Since all good types give the same in equilibrium, any point x_g is a fixed point of the continuous function $B(x_g) = b(\frac{g x_g}{N-1})$. By the Implicit Function Theorem applied to (3), $b' = -\frac{w_{12}}{w_{11}} > 0$. By our condition that $b'(X) = \frac{-w_{12}(x_g, X)}{w_{11}(x_g, X)} < k\frac{x_g}{X} = k\frac{N-1}{g}$, for $k < 1$, $B(\cdot)$ is a contraction on $[0, \bar{x}]$; also our conditions ensure that $B(x_g) \in [0, \bar{x}]$ for any x_g. Thus, B has a unique fixed point.

We define a symmetric equilibrium when good types are unknown, x^*, as a fixed point where $x^* = \beta(x^*)$. $x^* > 0$ since $\beta(0) > 0$, and it exists since $\beta(\bar{x}) \leq \bar{x}$ and β is continuous by the IFT. Implicitly differentiating (6) gives

$$\frac{d\beta(x)}{dx} = \frac{\sum \Pi(g) \frac{g}{N-1} w_{12}}{-\sum \Pi(g) w_{11}} > 0, \tag{7}$$

suppressing function arguments. By our condition on w_{12}/w_{11}, $w_{12}(\beta(x), \frac{g}{N-1}x) < -k(\beta(x)/\frac{g}{N-1}x)w_{11}$, so

$$\frac{\sum \Pi(g) \frac{g}{N-1} w_{12}}{-\sum \Pi(g) w_{11}} < \frac{\sum \Pi(g)(k\beta(x)/x) w_{11}}{\sum \Pi(g) w_{11}} \tag{8}$$

If $\beta(x) \leq x$, then $k\beta(x)/x < 1$ so the above is less than 1. Thus, if $\beta(x) \leq x$, then $\beta(x') < x'$ for $x' > x$. Therefore, $x^* = \beta(x^*)$ is unique. □

Proof of Lemma 1:

Common knowledge of the number of good types increases donations ex ante when $w_1(x, X)$ is weakly concave in X and $b(\cdot)$ is weakly convex.

Proof. Define $\bar{g} = \sum \Pi(g)(g+1)$ as the expected total number of good types. First, suppose that $f(g) \equiv (g+1)x_g$ is convex. Then

$$\sum_{g=0}^{N-1} \Pi(g)(g+1)x_g \geq \sum_{g=0}^{N-1} \Pi(g)(g+1)x_{\bar{g}}; \tag{9}$$

and if $x^* < x_{\bar{g}}$[34], (5) follows immediately, i.e. knowledge increases contributions.

We next prove that if the Lemma conditions hold, both the above conditions hold: $f(g)$ is convex and $x^* < x_{\bar{g}}$.

[34]The definition of x_g can be extended unchanged to non-integer values of g.

To show $x^* < x_{\bar{g}}$, first observe that

$$w_1(x^*, \frac{\bar{g}x^*}{N-1}) \geq E_g w_1(x^*, \frac{gx^*}{N-1}) = 1 - \bar{\alpha}, \tag{10}$$

the inequality by concavity of w_1, the equality by definition of x^*. Now suppose $x^* \geq x_{\bar{g}}$. Then $w_1(x^*, \frac{\bar{g}x^*}{N-1}) < 1 - \bar{\alpha}$, a contradiction. (Proof: $b'(\frac{\bar{g}x_{\bar{g}}}{N-1}) < \frac{N-1}{\bar{g}}$, as in the previous proof. So for some small ε and all $x \in (x_{\bar{g}}, x_{\bar{g}} + \varepsilon)$, $b(\frac{\bar{g}x}{N-1}) < x$. Suppose $b(\frac{\bar{g}x^*}{N-1}) \geq x^*$. Then at some point $y \in (x_{\bar{g}}, x^*]$, $b(\frac{\bar{g}y}{N-1}) = y$. But this would contradict uniqueness of $x_{\bar{g}}$. So $b(\frac{\bar{g}x^*}{N-1}) < x^*$. Then, since $w_{11} < 0$, we have $w_1(x^*, \frac{\bar{g}x^*}{N-1}) < w_1(b(\frac{\bar{g}x^*}{N-1}), \frac{\bar{g}x^*}{N-1}) = 1 - \alpha$.) Thus $x^* < x_{\bar{g}}$.

To show $f(g) = (g+1)x_g$ is convex, it suffices to show that x_g is convex. Now x_g solves $b(\frac{gx_g}{N-1}) - x_g = 0$. Applying the Implicit Function Theorem,

$$\frac{dx_g}{dg} = \frac{-\frac{x_g}{N-1} b'(\frac{gx_g}{N-1})}{\frac{g}{N-1} b'(\frac{gx_g}{N-1}) - 1}. \tag{11}$$

This is positive, by $b' > 0$ and $b' < \frac{N-1}{g}$, and we can rearrange it to

$$\frac{dx_g}{dg} = \frac{x_g/(N-1)}{\frac{1}{b'(gx_g/(N-1))} - \frac{g}{N-1}} > 0. \tag{12}$$

Now if g increases, then the top increases while the denominator decreases, since b' is weakly increasing by convexity of b. Thus dx_g/dg increases in g, showing that x_g is convex. □

The following Lemma is required for our proposition.

Lemma. *In the first stage, a separating equilibrium in which good types contribute $x < \beta(x)$ in the first round is not intuitive.*

Proof. In any separating equilibrium, bad types contribute less than good types, since otherwise they could increase their first round utility and simultaneously pool with good types, inducing greater contributions in the later round. Thus, bad types contribute $y < x$. Now say $x < \beta(x)$ and consider a deviation by a good type to $\beta(x)$. Since bad types prefer contributing y and being recognized as a bad type to contributing x and being recognized as a good type for sure, a fortiori they would not prefer to contribute $\beta(x) > x$ whatever the resulting belief. Good types, however, would prefer to contribute $\beta(x)$ than x, since $\beta(x)$ is a the good type's best response when other good types are contributing x and the distribution of good types is the prior. In particular, if the resulting belief is that the player is good for sure (or, in the anonymous case, that there is one more good type), then good types prefer to deviate to $\beta(x)$. But if so, good types have a credible deviation and the equilibrium is not intuitive. □

Proof of Proposition 2

Proposition 2, restated: In the anonymous first stage, there is an Intuitive separating equilibrium if and only if $D \geq \hat{D}$. In the revealed first stage, there is an Intuitive separating equilibrium if and only if $D \geq D^*$, where $D^* > \hat{D}$.

Proof. By the previous Lemma we can assume that good types contribute $x \geq \beta(x)$. In the revealed game, suppose there is a separating equilibrium where good types contribute $x \geq \beta(x)$ in the first round, bad types contribute 0. Beliefs are such that those who contribute x are believed good with 100% probability; those contributing less than x are believed bad with 100% probability; beliefs can be anything for those contributing more than x. These beliefs support play as specified, *if* there is separation in equilibrium: good types cannot do better than playing x, since $x \geq \beta(x)$ and their utility is concave, and bad types cannot do better than playing 0.

Good type donations, after g good types are revealed and only these players are included, are y_g satisfying $w_1(y_g, y_g) = 1 - \alpha/g$. Therefore, the bad type's incentive compatibility constraint (IC) to play 0 instead of x is

$$\sum_{g=0}^{N-1} \Pi(g) D\alpha \frac{g}{N} x \geq \sum_{g=0}^{N-1} \Pi(g) \left\{ D(\alpha \frac{g+1}{N} - 1)x + \alpha \frac{g}{g+1} y_{g+1} \right\}. \tag{13}$$

In the second round, if the bad type gives x, he will be included in a group of $g+1$, of whom g will give y_{g+1}. Simplifying this:

$$D(1 - \frac{\alpha}{N})x \geq \sum_{g=0}^{N-1} \Pi(g) \alpha \frac{g}{g+1} y_{g+1}. \tag{14}$$

For any x equilibrium we can now calculate the lowest D that satisfies the bad type IC. This will make the above hold with equality.

We show in Lemma 3 in the Online Appendix that when (14) is satisfied, the good type's IC is also satisfied. Therefore, the lowest D allowing for separation in the revealed institution will satisfy (14) with equality. The lowest D possible is when $x = 1$, giving:

$$D^* = \frac{\sum_{g=0}^{N-1} \Pi(g) \alpha \frac{g}{g+1} y_{g+1}}{1 - \alpha/N}. \tag{15}$$

Next we show that these separating equilibria are intuitive. For any $x \in [\beta(x), 1]$, if (14) holds with equality, the bad type is just indifferent between playing 0 and playing x. Thus, he would strictly prefer to play $y \in (0, x)$ if this would result in him being believed good for sure. Therefore, the good type has no credible deviation to $y < x$. The good type could credibly deviate to $y > x$ (since bad types would not do this for any resulting belief) but has no incentive to: since $x > \beta(x)$ and good type utility is concave, deviating to $y > x$ would reduce round 1 utility and could not improve on the belief the good

type induces by playing x. So far we have ensured that an equilibrium with $x \in [\beta(x), 1]$ and D such that (14) holds with equality is indeed intuitive. For even higher values of D, there is an equilibrium with $x = \beta(x) = x^*$ and (14) holding with strict inequality. Then good types have no incentive to deviate to any $y \neq x$, and bad types have no incentive to deviate to any $y > 0$. Thus, there is always an intuitive separating equilibrium for $D \geq D^*$.

Now, we turn to the anonymous institution and again seek conditions for a separating equilibrium. Thus, after g players are revealed as good types, all N players are included and good type donations are x_{g-1}. The bad type IC is

$$\sum_{g=0}^{N-1} \Pi(g) \left\{ D\frac{g}{N}\alpha x + \alpha \frac{g}{N} x_{g-1} \right\} \quad \geq \quad \sum_{g=0}^{N-1} \Pi(g) \left\{ D(\frac{g+1}{N}\alpha - 1)x + \alpha \frac{g}{N} x_g \right\}. \tag{16}$$

Rearranging, this becomes

$$D\left(1 - \frac{\alpha}{N}\right) x \geq \sum_{g=0}^{N-1} \Pi(g) \alpha \frac{g}{N} \left(x_g - x_{g-1}\right). \tag{17}$$

To show that for any x, the lowest D satisfying this will be less than the lowest D satisfying (14), it will suffice to show that $x_g \leq y_{g+1}$ for all g. Since $w_1(x_g, \frac{g}{N-1}x_g) = 1 - \alpha/N$, by the positive cross-partial we have $w_1(x_g, x_g) > 1 - \alpha/N \geq 1 - \alpha/(g+1) = w_1(y_{g+1}, y_{g+1})$. But then $x_g < y_{g+1}$. (Proof: by assumption, $\frac{w_{12}(x,x)}{w_{11}(x,x)} > -\frac{x}{x} = -1$, so that $\frac{d}{dx}\left(w_1(x,x)\right) = w_{12}(x,x) + w_{11}(x,x) < 0$.)

It remains only to prove that when the bad type's IC (17) is satisfied with equality, the good type IC is satisfied. This is shown in Lemma 4.

The arguments that the separating equilibria in the anonymous institution are intuitive closely parallel those for the revealed institution, and are omitted. □

Lemma 3. *In the revealed institution, the good type's incentive compatibility condition holds when the bad type's IC condition* (14) *holds with equality.*

Proof. The good type's IC is

$$\sum_{g=0}^{N-1} \Pi(g) D \left[\alpha \frac{gx + \beta(x)}{N} - \beta(x) + w(\beta(x), \frac{g}{N-1}x) \right] \leq \tag{18}$$
$$\sum_{g=0}^{N-1} \Pi(g) \left\{ D \left[\alpha \left(\frac{g+1}{N} - 1 \right) x + w(x, \frac{g}{N-1}x) \right] + \left[\alpha y_{g+1} - y_{g+1} + w(y_{g+1}, y_{g+1}) \right] \right\}.$$

Here, the left hand side is the benefit from playing the first round best response $\beta(x)$ rather than x, and thus being excluded in the second round. The right hand side is the benefit from playing x and being included in the second round with g other good types, whereupon everyone plays y_{g+1}. Simplifying this gives

$$\sum_{g=0}^{N-1}\Pi(g)D\left[(\frac{\alpha}{N}-1)(\beta(x)-x)+w(\beta(x),\frac{g}{N-1}x)-w(x,\frac{g}{N-1}x)\right]\leq \tag{19}$$
$$\sum_{g=0}^{N-1}\Pi(g)\left\{(\alpha-1)y_{g+1}+w(y_{g+1},y_{g+1})\right\}.$$

Will this be satisfied when the bad type IC just holds? Since $w_1>1$ and $x\geq\beta(x)$ the left hand side is less than $\sum_{g=0}^{N-1}\Pi(g)D(\frac{\alpha}{N}-1)(\beta(x)-x)\equiv D(1-\frac{\alpha}{N})(x-\beta(x))$. So the above will be satisfied if

$$D(1-\frac{\alpha}{N})(x-\beta(x))\leq\sum_{g=0}^{N-1}\Pi(g)\left\{(\alpha-1)y_{g+1}+w(y_{g+1},y_{g+1})\right\} \tag{20}$$

equivalently

$$D(1-\frac{\alpha}{N})x\leq D(1-\frac{\alpha}{N})\beta(x)+\sum_{g=0}^{N-1}\Pi(g)\left\{(\alpha-1)y_{g+1}+w(y_{g+1},y_{g+1})\right\}. \tag{21}$$

When the bad type IC just holds, we can replace the left hand side using (14), to give

$$\sum_{g=0}^{N-1}\Pi(g)\alpha\frac{g}{g+1}y_{g+1}\leq D(1-\frac{\alpha}{N})\beta(x)+\sum_{g=0}^{N-1}\Pi(g)\left\{(\alpha-1)y_{g+1}+w(y_{g+1},y_{g+1})\right\}. \tag{22}$$

Now,

$$w(y_{g+1},y_{g+1})=w(0,y_{g+1})+\int_0^{y_{g+1}}w_1(y,y_{g+1})dy>[1-\alpha/(g+1)]y_{g+1}, \tag{23}$$

by the FOC on y_{g+1} and concavity of w. So the right hand side is greater than

$$\sum_{g=0}^{N-1}\Pi(g)\left\{(\alpha-1)y_{g+1}+[1-\alpha/(g+1)]y_{g+1}\right\}=\sum_{g=0}^{N-1}\Pi(g)\alpha\frac{g}{g+1}y_{g+1} \tag{24}$$

and thus (22) holds with strict inequality. □

Lemma 4. *In the anonymous institution, the good type's incentive compatibility condition holds when the bad type's IC condition* (17) *holds with equality.*

Proof. The good type IC is

$$\sum_{g=0}^{N-1}\Pi(g)\left\{D\left[\alpha\frac{g+1}{N}x-x+w(x,\frac{g}{N-1}x)\right]+\alpha\frac{g+1}{N}x_g-x_g+w(x_g,\frac{g}{N-1}x_g)\right\}\geq \tag{25}$$
$$\sum_{g=0}^{N-1}\Pi(g)\left\{D\left[\alpha\left(\frac{g}{N}x+\frac{\beta(x)}{N}\right)-\beta(x)+w(\beta(x),\frac{g}{N-1}x)\right]+\alpha\left(\frac{g}{N}x_{g-1}+\hat{z}\right)-\hat{z}+w(\hat{z},\frac{g}{N-1}x_{g-1})\right\}$$

where $\hat{z}$ is a best response to g other good types who each contribute x_{g-1}; $\hat{z}=b(\frac{g}{N-1}x_{g-1})\in(b(\frac{g-1}{N-1}x_{g-1}),b(\frac{g}{N-1}x_g))\equiv$

(x_{g-1}, x_g) by increasingness of $b(\cdot)$. Rearranging, this becomes:

$$D\left[\left(1-\frac{\alpha}{N}\right)(x-\beta(x))+\sum_{g=0}^{N-1}\Pi(g)\left\{w(\beta(x),\frac{g}{N-1}x)-w(x,\frac{g}{N-1}x)\right\}\right]\leq \tag{26}$$
$$\sum_{g=0}^{N-1}\Pi(g)\left\{\alpha\frac{g}{N}(x_g-x_{g-1})-\left(1-\frac{\alpha}{N}\right)(x_g-\hat{z})+w(x_g,\frac{g}{N-1}x_g)-w(\hat{z},\frac{g}{N-1}x_{g-1})\right\}$$

and using $w(\beta(x),\frac{g}{N-1}x)-w(x,\frac{g}{N-1}x)\leq 0$ by $w_1>0$ and $\beta(x)\leq x$, the left hand side is less than

$$D\left(1-\frac{\alpha}{N}\right)(x-\beta(x))<D\left(1-\frac{\alpha}{N}\right)x. \tag{27}$$

On the right hand side, we can write

$$\begin{aligned}
& \textstyle\sum_{g=0}^{N-1}\Pi(g)\left[w(x_g,\frac{g}{N-1}x_g)-w(\hat{z},\frac{g}{N-1}x_{g-1})\right] && (28)\\
= \;& \textstyle\sum_{g=0}^{N-1}\Pi(g)\left[w(x_g,\frac{g}{N-1}x_g)-\psi(\hat{z},\frac{g}{N-1}x_g)+\psi(\hat{z},\frac{g}{N-1}x_g)-\psi(\hat{z},\frac{g}{N-1}x_{g-1})\right]\\
= \;& \textstyle\sum_{g=0}^{N-1}\Pi(g)\left[\int_{\hat{z}}^{x_g}\psi_1(\bar{z},\frac{g}{N-1}x_g)d\bar{z}+\psi(\hat{z},\frac{g}{N-1}x_g)-\psi(\hat{z},\frac{g}{N-1}x_{g-1})\right]\\
> \;& (x_g-\hat{z})\left(1-\tfrac{\alpha}{N}\right)
\end{aligned}$$

by the FOC for x_g , concavity of ψ and $\psi_2>0$. Plugging this inequality into (26) shows that the right hand side is greater than

$$\sum_{g=0}^{N-1}\Pi(g)\alpha\frac{g}{N}(x_g-x_{g-1}). \tag{29}$$

Putting this together with the bound (27) on the LHS of (26), we find that (26) holds with strict inequality so long as

$$D\left(1-\frac{\alpha}{N}\right)x\leq\sum_{g=0}^{N-1}\Pi(g)\alpha\frac{g}{N}(x_g-x_{g-1}) \tag{30}$$

which holds when the bad type IC condition (17) holds with equality. □

A.2 Robustness checks

Evidence for conditional cooperation interpretation: instrumental variables regressions

Table 8: IV regressions: determinants of followers' stage 2 contribution

	(1) IV: All Repetitions[1]	(2) IV: All Repetitions[2]
Avg (others') St.1 Contr.		-0.29
		(0.53)
Rvld. × Avg (others') St.1 Contr.		0.788
		(0.665)
Dummy: Rvld. Trtmt.	-1.740	-3.697
	(1.986)	(2.277)
Repetition	-0.044	-0.057
	(0.181)	(0.180)
Rvld × Rep.	0.178	0.187
	(0.219)	(0.223)
Avg. predn. for leaders	0.697**	0.745*
	(0.220)	(0.306)
Constant	1.093	1.637
	(1.766)	(1.768)
Observations	144	144
F-test: excluded instruments	23.08	13 .14
Hansen (J) test: p value	0.723	0.685
Endogeneity test : p value	0.386	0.600

Abbreviations: Rvld= Revealed, Contr.=Contribution, Trtmt=Treatment, St.=Stage, Predn.=Prediction
Robust (clustered by subject) standard errors in parentheses.
In anonymous treatments predictions were for (e.g.,) <the guy who contributed 4 ecus>.
[1] 2SLS estimation. Instruments: Average stage 1 contribution interacted with treatment,
... lags 1-4 of others' average stage 2 contribution in previous repetitions for subject, avg. of this over all previous repetitions.
[2] 2SLS estimation. Instruments: lags 1-4 of others' average stage 2 contr. in previous repetitions for subject,
... average of this over all previous repetitions.
+ p<0.10, * p<0.05, ** p<0.01

The excluded instruments in column 1 are the average stage 1 contribution (in that repetition) interacted with the treatment, and the first four lags (for that subject) of others' average stage 2 contributions, as well as the average of others stage 2 contributions over the subject's entire "history". Here identification relies on the assumption that, after controlling for a follower's beliefs (over leaders' average stage 2 contributions), leaders' stage 1 contributions have no *direct* effect on a follower contribution. Here the instrumental variables technique (2SLS) is used to deal with the possibility of time-varying correlated shocks to both a subject's beliefs and her generosity.

Column 2 approximately follows Smith (2012), and uses only the lag and "history" variable mentioned above as instruments for a subject's predictions. Both IV regressions support our interpretation, finding a significant and positive *effect* of a subjects' prediction on her stage 2 contribution. The instruments are very strong (see F-tests), and easily pass the Hansen test (J-Test) of over-identifying

restrictions.[35]

Evidence against embitterment driving differences in contributions between treatments

Table 9: Controlling for previous exclusion; repetitions 8-15, followers only

Dependent variable = Subject's stage 2 contr.		
	(1)	(2)
	OLS	Fixed-effects
Revealed	1.91	
	(1.76)	
Avg. (others') St. 1 Contr.	0.64	0.43
	(0.39)	(0.31)
Rvld × Avg. (others') St. 1 Contr.	-0.38	-0.74
	(0.53)	(0.51)
Repetition	-0.056	-0.13+
	(0.090)	(0.075)
Revealed×Repetition	-0.30**	-0.26*
	(0.11)	(0.11)
Prev. excluded	-1.27	-0.091
	(0.77)	(0.49)
Rvld. × Prev. excluded	1.74+	1.03
	(0.89)	(0.93)
Constant	3.58*	5.56**
	(1.37)	(0.79)
Observations	384	384
Sum coef.: Rvld. × excluded	.47	.94

+ p<0.10, * p<0.05, ** p<0.01

Robust (clustered by subject) standard errors in parentheses.

Abbreviations: St.=Stage, Rvld= Revealed, Rept.=Repetition, Prev.=Previously, Contr.=Contribution, Trtmt=Treatment

Table 9 gives the results of two regressions of second stage contribution for follower subjects in repetitions 8-15. We examine the later stages when behavior is were more likely to have converged, and when there is enough experience for embitterment to be a possibility. We focus on followers, to be consistent with Table 6.[36] The second column includes a subject-fixed effect, to control for the likelihood that stingier subjects are more likely to have been excluded. In each of these we include a dummy variable

[35]Note that subject fixed effects are no longer needed for unbiasedness if the instruments are valid; however, regressions with FE yield similar results (by request). We ran these final 2 columns with and without fixed effects, with various combinations of lag variables, and using the 2SLS and LIML techniques. The instrumented '*average guess*' coefficient is positive (between 0.5 and 1.7) and strongly significant across a variety of specifications, and the instruments are reasonably strong. As in Smith, the Hansen test (also known as Sargan and Basmann test) for the validity of the instruments fails to reject the null at the 10% level, but is close to (weak) statistical significance in some specifications.

[36]Still, our results for *"Sum coef: Rvld. × excluded*", "*Repetition × Rvld. Trtmt*", and *"Avg (others') St.1 Contr.*" are preserved qualitatively both (i) when we also include leader subjects, and (ii) when we include all 15 repetitions. As in table 6, there is no negative adjustment in row 2 when we include all repetitions. Tables available by request.

"prev. excluded" indicating whether a subject has been excluded in any previous repetition. The net coefficient on previous exclusion for the Revealed treatment ("Sum coef: Rvld. $\times$ excluded") is positive and insignificant in both columns. This offers evidence against an embitterment effect driving the relative patterns. Furthermore, even with these controls, stage 2 contributions decline faster in the Revealed treatment ("Repetition $\times$ Rvld. Trtmt." coefficients). Finally, the net effect of an greater average stage 1 contributions remains positive in the anonymous case and lower in the revealed case, although neither coefficient is statistically significant here.

B Online Appendix

Why does information help in this context?

In the introduction, we mentioned the evidence that "unenforceable" contributions tend to decline in VCM games with repetition, that the decline is greater when there is more feedback on others contributions (e.g., Chaudhuri and Paichayontvijit, 2011),[37] and the argument that this is driven by overoptimistic conditional cooperators learning to be disappointed about others' types. This would suggest that more information, as our Anonymous first stage seems to provide, should precipitate the decline in cooperation. Why then do we find the opposite? In our experiment, predicted contributions decline over time, but they decline significantly faster for the Revealed treatment, for which the first stages are fairly uninformative. This lack of information seems to lead to greater pessimism about others' play, a preference to cooperate less under this uncertainty, or both. This seems to be the main factor driving the greater decline in contributions; controlling for these beliefs, and for the "last repetition effect" (in Table 6, columns 1 and 7) , the differential trend term is fairly small.

For space concerns, we only present some key points. The literature is somewhat ambivalent about whether the decline comes from updated beliefs about types or from a coordination failure among conditional cooperators, or some combination of these. Chaudhuri and Paichayontvijit (2011) argue that the "heterogeneity in the initial distribution of beliefs among conditionally cooperative players" is critical to this dynamic. If some of these will always "put the wrong foot forward", this may have set off a cycle of misunderstanding and miscoordination.[38]

Furthermore, the conditional cooperation responses seem to be nonlinear, suggesting that it is not merely the average predicted contribution rate that matters, but its distribution. Chaudhuri and Paichayontvijit find that pessimists increase their contributions over time but "not enough to offset the sharp drop in contributions from the disillusioned optimists"; it is nonetheless possible that in another environment, the pattern could be reversed. Chaudhuri (2011) citing De Oliveira et al. (2009), argues that "... the mere presence of conditional cooperators... Is not enough, conditional cooperators need to know that there are no selfish types in their group for them to sustain cooperation." This apparent nonlinearity is particularly striking in Burlando and Guala (2005), where the perfectly sorted groups achieved nearly complete cooperation for the entire session."

Thus, while in a range of previous papers, information about others' actions seemed to cause a decline in average cooperation, there is no reason to assume it always must – our experiment is a

[37] Chaudhuri and Paichayontvijit (2011) use a partners protocol over 24 rounds with varying feedback and belief elicitation treatments. Their results contrast with Sass and Weimann (2012), who invite subjects back four times (with random termination) over one week intervals, with is no learning until the very end. These authors find a decline in conditional (elicited using the Fischbacher et al., 2001 version of the strategy method) and unconditional cooperation, with conditional cooperators becoming free riders. This argues against "learning about peers" as the *sole* cause of a decline.

[38] Chaudhuri and Paichayontvijit (2011) use a partners protocol over 24 rounds with varying feedback and belief elicitation by treatment. On the other hand, Sass and Weimann (2012) find a decline in conditional and unconditional cooperation in an environment without any feedback; the latter finding argues against "learning about peers" as the *sole* cause of a decline.

counterexample. Ermisch and Gambetta (2010) provide further support for this; they find that greater exposure to strangers (and weaker family ties) increases trusting behavior.

Subjects' Earnings, formal notation

Formally, the earnings of a subject i in a given repetition were:

$LEADER \times [(4 - S_i) + \frac{1.5}{3} \sum_{j=1}^{3} S_j]$ for Stage One, $10 + (1 - EXCLUDED_i) \times [(\frac{2}{M} \sum_{k=1}^{M} C_k) - C_i]$ for Stage Two. In addition, where i's prediction $\hat{C}_j^i$ of one other subject j's stage 2 contribution ($C_{j,}$) is chosen for a reward, the prediction reward is $20 - ((\hat{C}_j^i - C_j)^2/5)$.

Where $LEADER$ is a dummy indicating a leader subject, S_i is a player i's Stage-one contribution, $EXCLUDED_i$ is a dummy indicating that player i has been excluded, M is the number of non-excluded players in Stage 2 (note $M = 4$ if there has been an exclusion, $M = 5$ otherwise), C_k is player k's Stage-two contribution, , G_i is player i's guess. As noted, for each subject, a single repetition was randomly chosen for stage 1 payoffs, and a single repetition was randomly chosen for stage 2 payoffs. In each session, one guess from one guessing stage for a single subject is chosen for rewarding predictions.

Supplemental summary statistics

Table 10: Subject characteristics

Gender	Freq.	Percent	Cum.
Female	91	60.67	60.67
Male	59	39.33	100
Total	150	100	

Field of Study	Freq.	Percent	Cum.
Bioinformatics	1	0.68	0.68
Biology	6	4.08	4.76
Business Administration	14	9.52	14.29
Chemistry	1	0.68	14.97
Computer Science	3	2.04	17.01
Cultural Studies	2	1.36	18.37
Economic mathematics	2	1.36	19.73
Economics	3	2.04	21.77
Educational science	14	9.52	31.29
Engineering	9	6.12	37.41
English Language and Literature Studies	2	1.36	38.78
Geography	3	2.04	40.82
Geology	3	2.04	42.86
German Language and Literature Studies	10	6.8	49.66
History	4	2.72	52.38
House husband/Housewife	1	0.68	53.06
Law	18	12.24	65.31
Mathematics	4	2.72	68.03
Media science	2	1.36	69.39
Medical science	1	0.68	70.07
Musicology	1	0.68	70.75
Nutrition science	2	1.36	72.11
Pharmaceutics	2	1.36	73.47
Philology	2	1.36	74.83
Physics	3	2.04	76.87
Political Science	10	6.8	83.67
Psychology	2	1.36	85.03
Pupil	2	1.36	86.39
Slavic Languages and Literature	1	0.68	87.07
Sociology	13	8.84	95.92
Sports science	2	1.36	97.28
History	1	0.68	97.96
History of art	1	0.68	98.64
Public employee	1	0.68	99.32
Self-employed	1	0.68	100
Total	147	100	

University Entry Year	Freq.	Percent	Cum.
2002	7	4.76	4.76
2003	15	10.2	14.97
2004	16	10.88	25.85
2005	12	8.16	34.01
2006	22	14.97	48.98
2007	38	25.85	74.83
2008	30	20.41	95.24
don’t know	2	1.36	96.6
Not applicable	5	3.4	100
Total	147	100	

Prev. Experiments.	Freq.	Percent	Cum.
3-5	9	6	6
6-10	31	20.67	26.67
11-15	33	22	48.67
16-20	46	30.67	79.33
21-25	23	15.33	94.67
26-31	8	5.33	100
Total	150	100	

Votes to exclude as predictors of stage 2 contributions

Additional result: *Votes to exclude a subject were poor predictors of this subject's stage 2 contribution choice in the Revealed treatment. As a result, these targeted exclusions did not directly increase stage 2 payoffs for the remaining subjects.*

This is a natural consequence of the previous results; in the Revealed case, exclusion votes were strongly correlated to a leader's stage 1 contribution, but these stage 1 contributions were poor predictors of stage 2 contributions. This result is shown in Table 11, which reports regressions of stage 2 contribution behavior on the probability that a subject is excluded, plus controls. Columns 1 and 2 show a significant negative relationship between stage 2 contribution and the probability of exclusion in the Anonymous treatment only. In other words, in the Anonymous treatment, exclusions tend to be applied against (groups of) leaders who actually would contribute less in stage 2. In the first column the negative coefficient on "Prob. Excluded (simulated)" shows that excluded leaders tend to contribute less than non-excluded leaders. (The negative but insignificant coefficient in column 2 suggests that they tend to contribute less than all other non-excluded subjects, whether leaders or followers.) In contrast, as the small and insignificant summed coefficients show, exclusions in the Revealed treatment were ineffective at removing leaders who would contribute less in stage 2. To address the possibility that this different stage 2 behavior was driven by leaders' beliefs that they were likely to be excluded, in the final column we use the subject's average stage 2 contribution in *previous* repetitions as the dependent variable, finding similar results.

Table 11: Leaders' stage 2 contributions by conditional probability of exclusion

	(1) St.2 Contr.	(2) St.2 Contr.	(3) Avg. Prev. St.2 Contr.
Leader	0.85**		
	(0.18)		
Leader × Rvld Trtmt	-1.04**		
	(0.23)		
Prob. Excluded (simulated)[1]	-3.59**	-1.18	-2.15*
	(1.25)	(0.96)	(0.95)
Prob. Excluded × Rvld Trtmt	3.49**	0.96	2.00*
	(1.31)	(1.02)	(1.01)
Avg (others') St.1 Contr.	0.031	0.086	-0.24+
	(0.15)	(0.15)	(0.13)
Avg (others') St.1 Contr. (others) × Rvld Trtmt	0.27	0.23	0.40*
	(0.17)	(0.18)	(0.17)
Additional controls	Yes	Yes	Yes
Observations	1950	1950	1800
Sum coef.: Prob. excl. in rvld trtmt	-0.095	-0.22	-0.15

Abbreviations: St.=Stage, Prev. = Previous, Rvld= Revealed, Contr.=Contribution, Trtmt=Treatment, Excl.=Excluded

Robust (clustered by subject) standard errors in parentheses.

[1] Conditional expectation simulated for anonymous case; see notes after Table 1.

Additional controls: Session/treatment dummies, dummy: high example donation in instructions,

high example × revealed treatment, repetition, revealed × repetition, last repetition, revealed × last repetition.

+ p<0.10, * p<0.05, ** p<0.01

The effect of the instructional example [Available by request, not online, as we may use this as part of a subsequent paper]

Table 12: T-test and Wilcoxon rank sum test: effect of treatment by example

	Stage 1 contribution		Stage 2 contribution	
	T-test	Rank sum	T-test	Rank sum
Overall	-1.84+	0.427**	3.95**	0.631**
	(0.068)	(0.000)	(0.000)	(0.000)
High example	-0.03	0.497	2.02**	0.684**
	(0.976)	(0.881)	(0.00)	(0.000)
Low example	-2.67**	0.347**	1.37	0.568**
	(0.009)	(0.000)	(0.173)	(0.000)

Columns contain t-values and P(anon > revealed), respectively. P values in parentheses
T-tests from univariate regressions, on "Anon" dummy, std. errors clustered by id.
+ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$

Figure 6: Mean stage 1 contribution by repetition, treatment, and instructional example

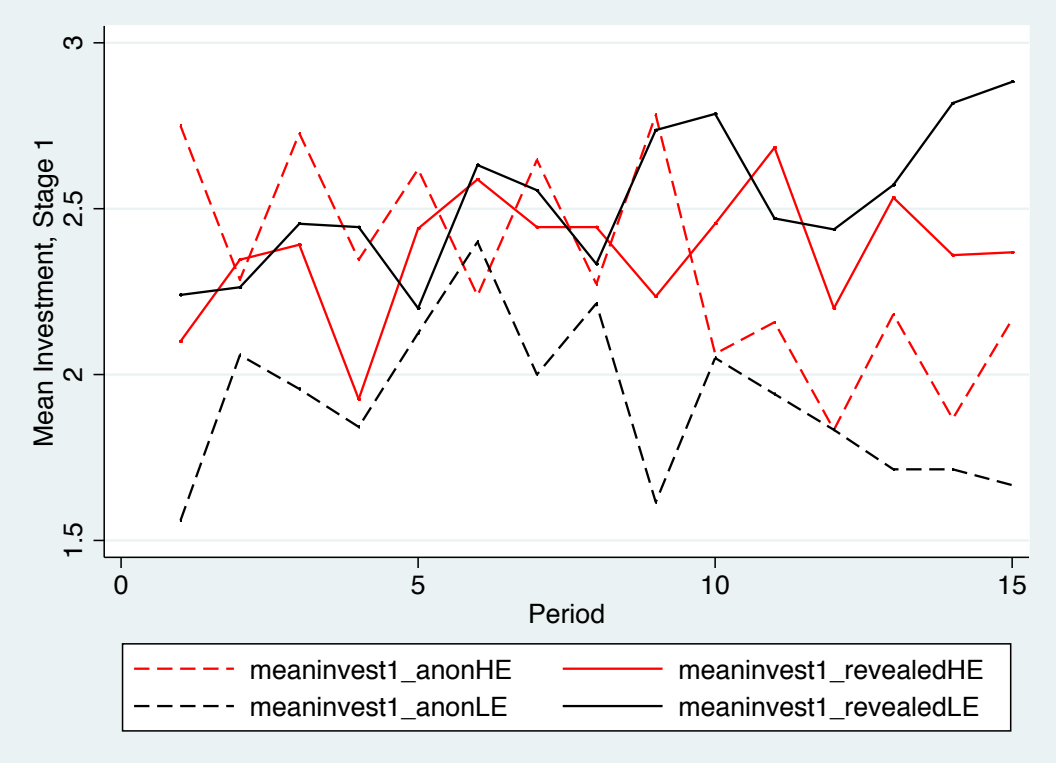
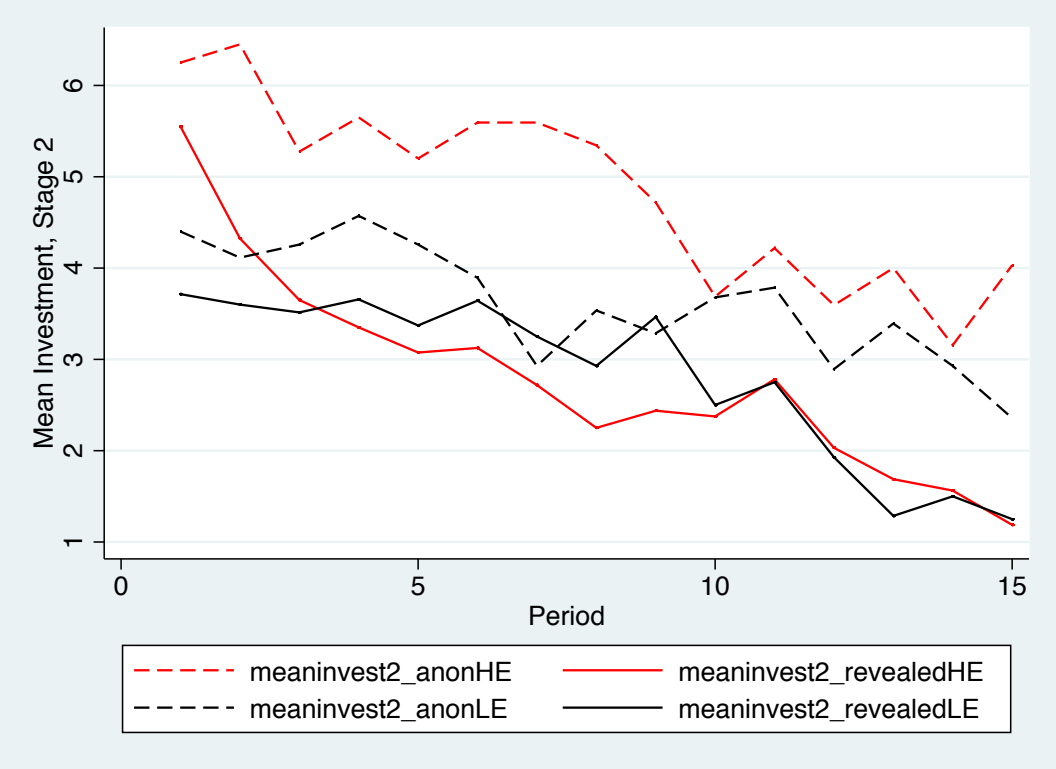


Figure 7: Mean stage 2 contribution by repetition, treatment, and instructional example



In above figures: Dashed = anonymous, Solid = Revealed; black = Low example, Red = High Example.

Instructions (English)

See file: "Instructions_(english) - 2 treatments x 2 examples.zip"

Screenshots

See file: screenshots.zip. (Translation by request).