Hugh-Jones, David; Reinstein, David

Working Paper

Secret santa: anonymity, signaling, and conditional cooperation

Jena Economic Research Papers, No. 2009,048

JENA ECONOMIC RESEARCH PAPERS

2009 – 048

Secret Santa: Anonymity, Signaling, and Conditional Cooperation

by

David Hugh-Jones
David Reinstein

www.jenecon.de

ISSN 1864-7057

The JENA ECONOMIC RESEARCH PAPERS is a joint publication of the Friedrich Schiller University and the Max Planck Institute of Economics, Jena, Germany. For editorial correspondence please contact markus.pasche@uni-jena.de.

Impressum:

Friedrich Schiller University Jena
Carl-Zeiss-Str. 3
D-07743 Jena
www.uni-jena.de

Max Planck Institute of Economics
Kahlaische Str. 10
D-07745 Jena
www.econ.mpg.de

Jena Economic Research Papers 2009 - 048

Secret Santa:

Anonymity, Signaling,and Conditional Cooperation

David Hugh-Jones and David Reinstein*

July 2, 2009

Abstract

Costly signaling of commitment to a group has been proposed as an explanation for participation in religion and ritual. But if the signal's cost is too small, freeriders will send the signal and behave selfishly later. Effective signaling may then be prohibitively costly. If the average level of signaling in a group is observable, but individual effort is not, then freeriders can behave selfishly without being detected, and group members will learn about the average level of commitment among the group. We develop a formal model, and give examples of institutions that enable anonymous signaling, including ritual, religion, music and dance, voting, charitable donations, and military institutions. We explore the value of anonymity in the laboratory with a repeated two-stage public goods game with exclusion. When first-stage contributions are *anonymous,* subjects are better at predicting second-stage behavior, and maintain a substantially higher level of cooperation.

Keywords: signaling, anonymity, public goods

JEL classification: H41

*David Hugh-Jones is a post-doctoral researcher at the Max Planck Institute for Economics, Jena. David Reinstein is a lecturer in the Department of Economics at Essex University.

Abstract

1 Introduction

Individuals in social settings often participate in joint activities that seem to yield no direct benefits to themselves or the group. In many of these activities, which we will broadly call "rituals," while each individual's effort is unobserved, some measure of the group's aggregate effort can be seen by all the participants, and often by outsiders as well. We suggest that in some cases, such anonymous rituals evolve not to signal individuals' type or fitness, nor to screen out "bad" types, but to *measure* the level of cooperativeness in the group as a whole. This knowledge will allow the less cooperative groups to dissolve, and embolden the more cooperative groups to undertake important collective tasks which are vulnerable to shirking and free-riding. Anonymity lowers free-riders' incentives to pool with cooperators, since they cannot be singled out for punishment or exclusion. We offer three motivational examples.

1. A charity is raising money for a good cause. The project will only succeed with enough commitment from donors and volunteers, and since time and money are costly, each potential donor needs to be persuaded that other donors are "on board". Early donations of "seed money" could accomplish this. But named donations may be given only to win credit and publicity for the giver, and may not indicate real long-term commitment. Anonymous donations bring no reputational benefit, so they are better signs of real commitment and elicit higher gifts from others.

2. A group of co-workers wonder if they can strike for higher wages. Will they support each other, or will some blackleg their colleagues? To work out the level of mutual goodwill, they think back to the Christmas presents they received from their workmates. But these may have been given in order to strike up mutually beneficial long-term relationships. On the other hand, if the workplace has the "secret Santa"

institution, in which present-givers are anonymous to the receivers, then the generosity of gifts conveys real information about co-workers' character.

3. Members of a church wish to do business with each other. They would like to know whether their co-religionists are more honest than outsiders. If contributions to the church are made anonymously and voluntarily, the proportion of their fellow congregants who tithe may be a good indicator. On the other hand, if contributions are made publicly, families may be tithing only to assimilate into the community (and possibly defraud them later).

In all of these examples, participants can anonymously perform an altruistic action. Other participants and observers then learn about the level of altruism in the society or the group. This learning may be useful in future collective action problems in which the stakes are higher, as it allows "conditional cooperators" (see Fischbacher et al. (2001), Schram (2000)) to contribute only when they are certain enough that their actions will be reciprocated.

This argument requires there to be different types of individual: selfish actors and conditional cooperators. Selfish actors prefer to shirk in the main collective action problem. Conditional cooperators prefer to contribute, but only if contributions from other players are sufficiently high.[1]

The next section discusses the existing literature on costly signaling, and shows the contribution of this paper. Section 3 gives more in-depth examples of institutions to facilitate anonymous signaling. We then develop a formal model. Section 5 presents our laboratory experiment testing the model and its predictions. The conclusion discusses the model's implications.

[1]There might also be altruistic, or "unconditionally cooperative" types. If anonymous signaling reveals enough of these, cooperation may be a dominant strategy for conditional cooperators.

2 Literature

The problem of collective action is well known and has been studied extensively across the social and biological sciences. While cooperation may be assured by externally imposed sanctions, or by an equilibrium in a repeated game that acts as a sanctioning mechanism (Fudenberg and Tirole 1991), many situations do not allow these solutions. Surprisingly, cooperation is nevertheless often observed. A recent set of theories involve heterogeneous preferences: some players are purely self-interested while others are reciprocators who will cooperate if they expect enough others to do so (Kreps et al. 2001, Schram 2000, Ostrom 2000). In fact, experimental economists have found considerable evidence for "conditional cooperation" - see Fischbacher et al. (2001) - and also for heterogeneity (e.g., Simpson and Willer 2008).[2]

In this environment, players have an interest in learning whether their companions are selfish or reciprocal, since this may determine how much they themselves are willing to cooperate. One way to find out is by observing cooperation on a small scale first. Reciprocators can then be separated from selfish players by their willingness to take part. This could be true even if the "cooperative" behaviour is wholly symbolic. For instance, ritual and religion have been explained as "costly signaling" ways for individuals to show their commitment to a particular group (Weber 1946, Ruffle and Sosis 2003, Sosis and Ruffle 2003, Bird and Smith 2005, Levy and Razin 2006).

Of course, if one's behaviour in a small-scale or symbolic setting may affect others' cooperation in an important setting, even the self-interested have strong incentives to "play nice". In particular this will hold if others are motivated to exclude selfish types from the collective good, whether to avoid crowding or out of a desire to punish (Ostrom et al. 1992, Fehr and Gachter 2000). So cooperation in the small-scale setting must be costly enough to deter selfish players from acting like reciprocators. But then the signaling institution will

[2]This theory is not essential for our model. Differences in type need not reflect differences in underlying motivations, but could come from differences in individuals' material benefits from a public good or differences in their options outside the group.

impose a large cost on those who take part.[3]

Therefore, groups may face a commitment problem. All would be better off *ex ante* if those who behaved selfishly in the small-scale setting did not face sanctions, since this would allow types to be separated at lower social cost; but *ex post* it will be tempting to sanction or exclude selfish individuals. We propose that societies can avoid this problem by using an *anonymizing technology* that reveals only the *overall* level of cooperative behaviour in a smaller-stakes setting. Thus, the number of cooperative players can be observed accurately at a lower cost, since there is less incentive for selfish players to pool.

The existing literature on social dilemmas mainly treats anonymity as a problem to be solved. Game theorists have shown the possibility of cooperation in the repeated Prisoner's Dilemma when play is anonymous, enforced by a "punishment equilibrium" in which everyone defects (Ellison 1994, Kandori 1992). Experiments have found that being identified increases people's donations to charity under some circumstances (Carman 2003, Soetevent 2005, Andreoni and Petrie 2004a, Reinstein and Riener 2009). In the present paper, by contrast, anonymity can be beneficial.

The advantages of anonymity have been discussed in the literature on transparency in principal-agent relationships. When only incomplete contracts are possible, and agents have "career concerns", principals may benefit from committing not to learn too much about the choices made by agents (Holmstrom 1999). Being observed may lead agents to choose an action which makes them look smart, rather than the one which is actually best for the principal. In a political context, Prat (2005) and Levy (2007b, 2007a) show that the secret (i.e., anonymous) ballot may improve legislative outcomes. Acemoglu (2007) models firms' commitment problem when they are *ex ante* better off not examining individual employees' performance.

To sum up, this paper makes three contributions to the existing literature. First, it shows that conditional cooperators engaged in collective action, under conditions of uncertainty,

[3] An extreme example of the cost of ritual behaviour is provided by Saint Simeon Stylites, who lived for 37 years on top of a pillar. The practice continued for centuries after his death.

face a problem of "pandering" which makes it hard for them to know whether they can trust each other. Second, it proposes that anonymity mitigates this problem problem of pandering, and therefore has a previously-ignored positive role in developing trust. Third, it suggests that certain institutions - in particular, some collective rituals and "symbolic" collective actions - may have developed to provide an anonymous contribution technology, which allows insiders and outsiders to cheaply learn the *real* average level of cooperativeness in the group. Finally, we demonstrate this with a laboratory public goods game in which anonymity increases contribution levels.

3 Examples of Anonymous Signaling

Many charities identify and thank donors who give a specified amount. These doubtless elicit donations by people in search of social recognition. Hence, Andreoni and Petrie (2004b) mention it as a puzzle that charities allow anonymous donations. However, in many settings anonymity is the norm, and only total donations are made public, without identifying individual givers. Although this may lower giving by reputation-seekers, it may serve other purposes. A high rate of anonymous giving may lead to an environment of mutual trust. Furthermore, early anonymous donations may possess greater signaling value in encouraging others to come forward.[4] This is especially true when there is a small community of potential givers, who are known to one another, so that non-anonymous donations have a strong effect on reputation. Notably, church collections are often taken in a pocket that conceals the amount of individual donations, or via anonymous checks. Whatever the avowed reason, our story illustrates how anonymous donations can make the total amount collected more informative about church members' community spirit.

The need to preserve anonymity may favour forms of ritual behaviour which are of little practical use. Dance has long been recognized as a way of enhancing group solidarity, and it

[4] List and Lucking-Reiley (2002), List and Rondeau (2003), Carman (2003), Karlan and List (2007), Reinstein and Riener (2009) provide some evidence on the signaling value of anonymous and identified leadership contribution.

has even been suggested that the development of this function was an essential step in human evolution (McNeill 1997). Hagen and Bryant (2003) discuss song and music as an example of "coalitional signaling". In certain forms of music and dance, individuals' levels of effort are masked while the average level is obvious. Group singing can be judged by its overall volume, harmony and enthusiasm, while it is harder to discover exactly who is out of tune or keeping quiet. In communal dances, if one person makes a mistake, the entire group may lose the rhythm, and the guilty party can not always be identified.

Armies need to signal, to themselves and their opponents, that their members are truly committed to risking their own lives to achieve victory. Military music has a long tradition. When infantry march in step, a single person who loses the rhythm will cause the whole unit to break step. This is an anonymous, weakest-link signaling technology. Vegetius ([c. 400] 2004) emphasizes the importance of disciplined marching in his handbook for the training of recruits to the Roman legions: "troops who march in an irregular and disorderly manner are always in great danger of being defeated."

Modern armies also use uniforms. A large literature in sociology and social psychology examines the effects of uniform on behaviour. A key idea is that uniforms "deindividuate" their wearers, making them more likely to conform to group norms (Rafaeli and Pratt 1993, Joseph and Alex 1972). Watson (1973) found that cultures which used warpaint or other anonymizing techniques were more likely to fight to the death and take no prisoners. The deindividuation may originate from the sense that one has become anonymous. Anonymous individuals cannot use the battlefield as a stage to display their courage; this means that an opponent who seems aggressive probably really is so, and will not back down if challenged.

In a modern context, we may be able to judge the culture of an organization or group by small-scale institutions like the aforementioned Secret Santa. We can observe the size of donations to the "honesty box" for office coffee, or the quality of the food brought to crowded "pot luck" parties. When it is time to pay a bill at the restaurant and everyone contributes "what they think they owe" does the sum exceed the bill or does it come up short?

Applause, cheers and jeers are all reliable signals of collective appreciation, since contribution levels cannot be distinguished. An anecdote from Solzehnitzyn (1997) shows what happens when anonymity breaks down: in a Soviet conference, applause for Stalin lasted eleven minutes, and the first man to stop clapping was singled out and arrested.

Since the nineteenth century the ballot in Britain and the US has been secret. For political parties, this had the disadvantage that voters could no longer be bribed into supporting them. However, there was also an advantage: a party's vote share became a reliable signal of public support for its policies. As one function of voting is to signal one's support, both to the party and to other social actors (Smirnov and Fowler 2007, Stigler 1972, Londregan and Vindigni 2006), political parties may have gained by making the signal clearer.

The disadvantage of keeping individuals' effort levels anonymous is that some information is lost. Observers learn only about the overall level of contributions to a public good, not about who contributes. When a symbolic collective display is meant to inform people outside the group, this disadvantage no longer matters. For example, if a war chant is meant to impress an opposing group with the readiness of participants to fight, the opponents will want to know how many fanatical warriors they are facing, while the singing group has no incentive to signal exactly who they are. Thus, although we do not model this here, anonymity-preserving institutions appear to be especially well-suited to signaling to outsiders.

4 Model

N agents each own $1/N$ of a unit of wealth. They may contribute it to a collective good,[5] and individual i's resulting utility is

$$w^{\tau}(x_i, X_{-i}, P)$$

[5]The good has both private and public elements. For some types benefit will be a function both of own contribution and aggregate contributions. We will also introduce rivalry (crowding) and exclusion. Still, the full benefits of the good will not necessarily be internalized.

where $x_i \in [0,1]$ is the share of his wealth he contributes, $X_{-i} = \sum_{j \neq i} x_i / N$ is the total wealth contributed by others, $P \in [0,1]$ is the proportion of agents included in the collective good, and $\tau \in \{G, E\}$ is i's type which may be conditionally cooperative (pro-social) or anti-social (selfish), or more jocularly good (G) and bad/evil (E). We will mostly be dealing with the good type's welfare w^G so we notate this as $w \equiv w^G$ for convenience. Initially, we will not allow exclusion, and therefore write

$$w^\tau(x, X) \equiv w^\tau(x, X, 1)$$

for welfare when no agent is excluded.

We can impose different restrictions on the form of w and w^E. Our core assumptions are as follows, where w^τ is short for "both w and w^E".

1. $w^\tau(x, X, P)$ is strictly increasing in X: other people's contributions always increase welfare.

2. $w^\tau(x, X, P)$ is strictly decreasing in P for $X > 0$: crowding lowers welfare.

3. $w(x, X, P)$ has increasing differences in x and X. That is, for all P, for $X' > X$ and $x' > x$, $w(x', X', P) - w(x, X', P) \geq w(x', X, P) - w(x, X, P)$. In economic terms, other people's contributions are complements to a good type's contributions. This is the most general formal expression of the idea of reciprocal altruism. It might also hold for strictly material reasons, for example, if the collective good has a weakest-link technology, then one's own contribution will not increase welfare unless others are contributing at least as much.

4. $x > 0$ is a maximizer of $w(x, X)$ for at least some X. Without this assumption, the collective action problem is trivial: nobody ever contributes. Combined with the previous assumption, this ensures that $w(\cdot, X)$ has positive maximizers for all X in some interval $[\underline{X}, 1]$.

5. $w^E(0,X,P) \geq w^E(x,X,P)$ for all x, X and P. Thus, evil types never contribute.

6. For all x, all $X' > X$ and all P, $w(x,X',P) - w(x,X,P) > w^E(x,X',P) - w^E(x,X,P)$. This is a crucial substantive assumption: not only are evil types less interested in contributing themselves, they are also affected less than good types by others' contributions. This allows signaling to separate the types.[6]

7. Finally we assume $w^\tau(0,0,P) = 0$ for all P.

As we shall see, these very general conditions are already enough to allow the existence of signaling institutions. Sometimes, to ensure unique interior equilibria, we add another assumption:

8. w^τ is differentiable. In this case, Assumption 3 means that $w_{12}(x,X,P) \geq 0$, and Assumption 6 means that $w_2(x,X,P) > w_2^E(x,X,P)$ for all x,X and P. (Subscripts denote the derivative with respect to the corresponding argument(s).) Furthermore, w is strictly concave, $w_1(0,0) > 0$ and $w_1(1,1) < 0$. Then Assumption 4 implies $w_1(x,X) \geq 0$ for some x and X, while Assumption 3 implies $w_1(1,X) < 0$ for all X. Finally, define $z(\cdot)$ implicitly on the appropriate subset of $[0,1]$ by the equation $w_1(x,z(x)) = 0$, i.e., by the level curve of w_1 at zero. By Assumption 3 , $z(x)$ is increasing.[7] To ensure uniqueness we assume that $z'(x) > 1$.

A type τ agent's prior belief is that there are g other good types with probability $P^\tau(g)$, $g \in \{0,...,N-1\}$. Sometimes we wish to increase N while holding the structure of the problem constant. We assume then for large N, $P^G \approx P^E$, i.e., one's own type is not very informative about others' types, and that the *proportion* of good types is distributed with continuous cdf $F(\gamma)$, supported on $\gamma \in [0,1]$.[8]

[6] Preferences are invariant to positive affine transformations in utility. However, we will shortly introduce other elements into agents' overall utility, so this assumption will be substantively meaningful.

[7] $z(x)$ could be a correspondence if $w_{12}(x,X) = 0$ for certain values of x.

[8] If types are drawn independently, then γ will become almost certain as N becomes large. But types need not be independent. For example, people's preferences may be affected by those around them, or by collective culture or education, whose effectiveness is not perfectly observable.

4.1 Equilibrium without knowledge

If X is known, players solve

$$\max_{x\in[0,1]} w^{\tau}(x,X) \tag{1}$$

and evil types never contribute anything. If the value of X is risky with cdf Φ (continuous or not) then evil types still never contribute anything, and good types solve

$$\max_{x\in[0,1]} \int w(x,X)d\Phi(X). \tag{2}$$

Write $b(X)$ or $b(\Phi)$ for the best response correspondences in the certain and risky cases. Suppose that these are single-valued. Then any equilibrium must be symmetric between the good types. For, suppose $x_i < x_j$ with both i and j good. Then $\sum_{k\neq i} x_k > \sum_{k\neq j} x_k$, for any possible contributions by other players. But then, Assumption 3 gives that $b(\sum_{k\neq i} x_k) > b(\sum_{k\neq i} x_k)$, so that x_i and x_j cannot be best responses. This also holds in the risky case since the increasing differences property is carried over to the integral in (2).

In a symmetric equilibrium when x^* is contributed by good types, $X = gx^*/N$ for a good type, where g is the number of *other* good types. The expected value to a good type of an interior equilibrium is

$$\sum_{g=0}^{N-1} P^G(g)w(x^*,gx^*) \tag{3}$$

and this is positive since $w(0,0) = 0 < w(0,gx^*) \leq w(x^*,gx^*)$ by Assumption 1 and optimality.

Suppose Assumption 8 holds. Then good types' best responses are single-valued and interior, with $w_1(b(X),X) = 0$, or in the stochastic case, $\int w_1(b(\Phi),X)d\Phi(X) = 0$. An equilibrium has $x^* \in (0,1)$ satisfying

$$\sum_{g=0}^{N-1} P^G(g)w_1(x^*,gx^*) = 0. \tag{4}$$

Assumption 8 does not guarantee uniqueness of x^*, since it makes no assumptions about

exactly how $w_1(x,X)$ varies with X. However, all our proofs hold for any value of x^*, so this is not a major problem.

4.2 Equilibrium with knowledge: when does complete information help?

Suppose that the total number of good types is known to be $g+1$. Then when good type contributions are x, other players' contributions will sum to $X = gx/N$. Equilibrium then has x satisfying (1) for this value of X. Write x_g for the largest such value of x: the equilibrium contribution of a good type when there are g *other* good types. By monotone comparative statics (Assumption 3), x_g is increasing in g (Milgrom and Roberts 1990). We assume that this equilibrium (which is Pareto optimal by $w^\tau(x,X)$ increasing in X for $\tau \in \{g,e\}$, and by optimality of x_g for good types) is selected. This is a reasonable assumption in many contexts, for instance if our agents are able to communicate.[9]

The value of an equilibrium to a good type when there are g *other* good types is:

$$V(g) = w(x_g, gx_g/N)$$

The good type's expected value is then

$$\sum_{g=0}^{N-1} P^G(g)V(g). \tag{5}$$

If w is differentiable and concave, as in Assumption 8, x_g satisfies

$$w_1(x_g, gx_g/N) \begin{cases} \leq 0 & \text{if } x = 0 \\ = 0 & \text{if } x \in (0,1) \\ \geq 0 & \text{if } x = 1 \end{cases} \cdot \tag{6}$$

[9] In addition to the theoretical justification (e.g., Farrell and Rabin (1996)) there is a great deal of experimental evidence for this. For example, in Cooper et al. (1994) 2-sided communication yields coordination on the efficient equilibrium in a stag-hunt type game 90% of the time.

This is illustrated in Figure 1. As this shows, there may be multiple and non-interior equilibria. The solid line in the picture shows the level curve $z(x)$ satisfying $w_1(x,z(x)) = 0$. This is increasing by Assumption 3. Then interior equilibria exist at points where the straight line $X = gx/N$ crosses $X = z(x)$. Equilibria also exist at $x = X = 0$ (if $w_1(0,0) \leq 0$), as in figure 1), and at $x = 1, X = g/N$ (if $w_1(1, \frac{g}{N}) \geq 0$. The equilibrium at zero is stable so long as $w_1(0,0)$ is strictly less than 0, and similarly the equilibrium at $x = 1$ is stable if $w_1(1, \frac{g}{N}) > 0$. Interior equilibria where $z(x)$ crosses gx/N from above are unstable, since a small shift up (down) along the gx/N line would make $w_1(x, gx/N) > 0$ (< 0) and so make it rational to contribute more (less). Interior equilibria are stable if $z(x)$ crosses $X = gx/N$ from below. In general the largest equilibrium at x_g is stable: either the level curve crosses $X = gx/N$ from below at an interior x_g, or $x_g = 1$.

The other parts of Assumption 8 ensure a unique interior equilibrium. Since $w_1(0,0) > 0$ and $w_1(1,X) < 0$ for all X, there can be no boundary equilibria. Since $z(x) - gx$ is increasing in x for all g, by $z'(x) > 1$, the interior equilibrium is unique.

We would like to know when (5) is an improvement on (3), i.e. when it is better for the number of good types to be publicly known. It is not always. For example, suppose that $w(x,X) = \tilde{w}(x+X)$ where $\tilde{w}$ is convex. Then good types will contribute either all or none of their endowment. If the distribution of the number of good types is just high enough to allow full contributions, then knowledge will always make things worse since for g low enough, good types will prefer to contribute nothing. The welfare comparison relies heavily on how $w(x,X)$ changes with total contributions by others, X. Since under Assumption 8, equilibrium conditions can be specified in terms of w_1, two w functions can have the same equilibria but different welfare. We use this to give a simple sufficient condition for knowledge of g to be beneficial: roughly, whenever good types' own contributions matter enough to their welfare.

Lemma 1. *Fix w^τ satisfying Assumption 8, $\tau \in \{G,E\}$, and set $w^{i\tau}(x,X) = w^\tau(z^{-1}(X),X) + i[w^\tau(x,X) - w^\tau(z^{-1}(X),X)]$ for $i > 0$. Then the $w^{\tau i}$'s satisfy our assumptions including Assumption 8, and for i high enough, good type welfare is higher when g is known than when g*

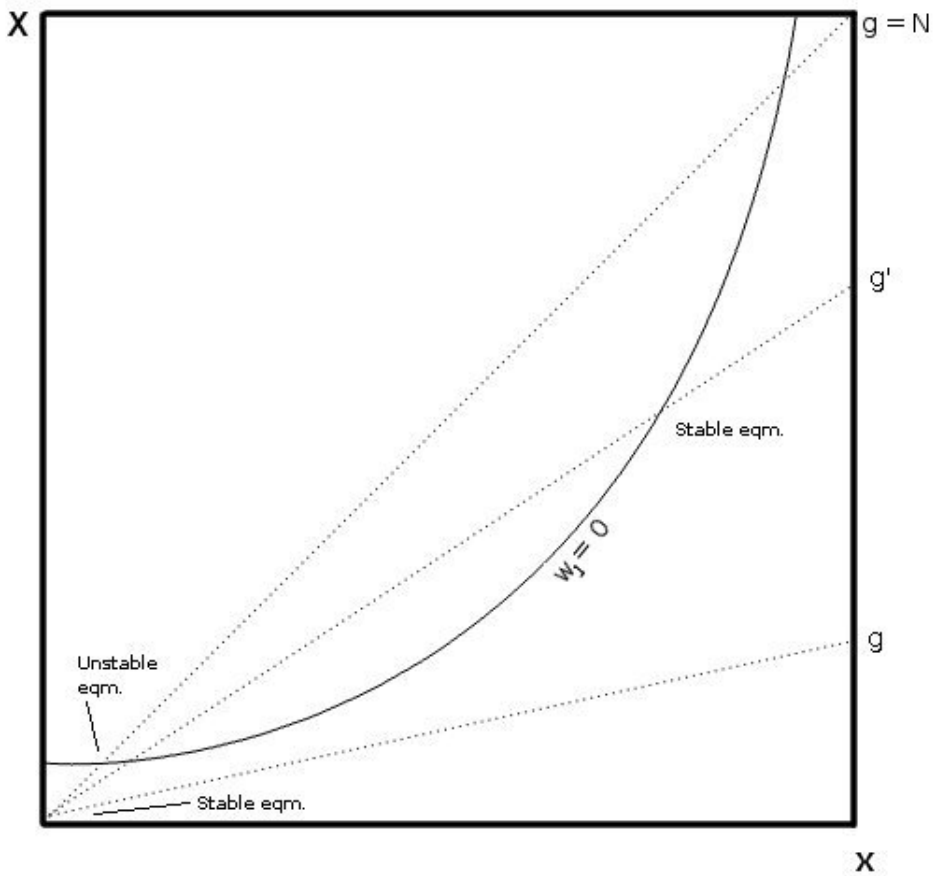


Figure 1: Equilibria when g is known

is unknown.

Proof. See the Appendix. □

The intuition is straightforward: knowing g enables good types to condition on others' behaviour better. The downside is that other players may do the same, which may lower total contributions; when own contributions matter a lot, the first effect dominates. In the Appendix, we also give a set of conditions for knowing g to be welfare-improving, which do not rely on $w(x,X)$ changing steeply in x.

4.3 signaling institutions

Cheap talk will not allow types to separate; since x_g is increasing in g, $w^E(0,kx_g)$is also increasing in g for all k, so that evil types have an incentive to claim they are good.

Suppose instead that either type can pay a cost $\sigma \geq 0$ before the game, and this payment is publicly observed. We focus on conditions for a separating equilibrium. Under a full pooling equilibrium, signaling institutions are either irrelevant or simply impose a deadweight cost on all agents. An equilibrium which separates the types always has only good types paying the cost. Otherwise bad types could strictly increase their welfare by not paying the cost, thus increasing others' prediction for g and increasing x_g.

For good types to pay in such a separating equilibrium, we must have

$$\sum_{g=0}^{N-1} P^G(g)V(g) - \sigma \quad \geq \quad \sum_{g=0}^{N-1} P^G(g)\left\{w(\hat{x_g}, gx_{g-1}/N)\right\} \tag{7}$$

where $\hat{x_g}$ maximizes welfare given total contributions of gx_{g-1}/N, i.e., when the g other good players think there are only g good players in total because of our agent's non-payment. The left hand side is just expected welfare from a signaling equilibrium, minus the signaling cost. Rearranging:

$$\sigma \leq \sum_{g=0}^{N-1} P^G(g)\left\{w(x_g, gx_g/N) - w(\hat{x_g}, gx_{g-1}/N)\right\} \tag{8}$$

Now $w(\hat{x_g}, gx_{g-1}/N) \leq w(\hat{x_g}, gx_g/N) \leq w(x_g, gx_g/N) \equiv V(g)$, the first inequality by w increasing in its second argument and x_g increasing in g, the second by optimality of x_g. [10]Therefore, there is a $\sigma \geq 0$ satisfying the good types' incentive compatibility constraint. Intuitively, good types discourage others from contributing if they pretend to be bad, so they will be prepared to pay some amount to send an honest signal. The value to a bad type when there are g good types in total (so that each good type observes $g-1$ other good types, and

[10]And when x_g is strictly increasing in g, for instance when it is interior and $w_{12} > 0$, the inequality will be strict.

plays x_{g-1}) is $w^E(0, gx_{g-1})$. Using this, the equivalent to (8) for bad types not to pay is

$$\sigma \geq \sum_{g=0}^{N-1} P^E(g) \left\{ w^E(0, gx_g/N) - w^E(0, gx_{g-1}/N) \right\} \tag{9}$$

and this is clearly a non-negative lower bound, by similar reasoning. A separating equilibrium is possible if (8) and (9) can hold simultaneously, and the minimum signaling cost will satisfy (9) with equality.

Lemma 2. *If $P^E(g) \approx P^G(g)$, or if $w^E(0,X)$ is differentiable with $w_2^E(0,X) < \varepsilon$ for low enough ε, then a separating equilibrium is possible. Furthermore if $w_2^E(0,X) < \varepsilon$ or N is large, the minimum signaling cost approaches 0.*

Proof. See the Appendix. □

The intuition behind this lemma is rather simple. A bad type can increase other players' contributions by pretending to be a good type, at a cost of σ. But good types care more about the collective good than bad types. So the maximum cost they will pay to increase others' contributions is higher than the minimum cost which will deter the bad type.[11] Furthermore, when N is large, any individual's contribution will make little difference to the total and therefore have little effect on others' behaviour. (Recall that total wealth remains unchanged at 1.) Also, the chance of being the "marginal" type who causes a big jump in equilibrium donations will be small. Hence the cost of signaling need not be high to deter bad types. The condition that $P^G \approx P^E$ will also hold for small N if agents' types are approximately independent.

When the conditions for $\sigma \to 0$ hold, and complete information offers a welfare improvement over equilibrium without information, then signaling will also offer this welfare

[11] If $P^E(g)$ differs substantially from $P^G(g)$, the result may not hold. For example, let $N = 3$. There is 1 good type with probability 1/2, and 0, 2 or 3 good types with probability 1/6 each. If there are at least two good types, all good types are prepared to contribute $x = 1$. Otherwise nobody will contribute anything. Now, in a separating equilibrium a good type who pays σ will only be decisive when there is 1 other good type, which occurs with probability $P^G(1) = 1/5$ by Bayes' rule. A bad type will influence contributions with a 3/5 probability, $P^E(1) = 3/5$. Then if the relevant payoffs of the two types, $w(1,2)$ and $w^E(0,1)$, are not too different, the bad type will be prepared to pay more than the good type, and no separating equilibrium will be possible.

improvement, since it allows learning about the types in the population at low cost.

4.4 Exclusion and crowding

Now suppose that after the signaling mechanism, but before the basic game, it is possible to exclude some or all players. With this addition crowding becomes relevant; recall that welfare is $w^{\tau}(x,X,P)$ where P is the proportion of players included, with w^{τ} decreasing in P.

There are many possible exclusion mechanisms that capture our intuition. The simplest is that a player is chosen at random to exclude or include all players. All players then wish to include themselves, but when N is large we can ignore this; otherwise, all players will include only good types and only up to the point that their equilibrium contributions outweigh the crowding effect. Alternatively, inclusion might be decided on a case-by-case basis by a majority vote. In this case, if there is a separating equilibrium, there will be unanimous agreement to include every good type, so long as their expected contributions outweigh crowding, and a $N-1$ against 1 vote to exclude every bad type, since even bad types wish to exclude other bad types. For simplicity we will assume that, as in these examples, the exclusion mechanism maximizes good types' welfare given the information available (the outcome of the signaling institution).

We consider two kinds of signaling institution; the difference is only relevant if exclusion is possible (otherwise both lead to the results described in the previous section). In a *public* signaling institution, as before, each agent's choice to pay σ or not is visible to all agents before exclusion takes place. In an *anonymous* signaling institution, only the number of agents who paid σ is visible.

The incentive constraints in a public signaling institution are now considerably tougher, as refusal to pay the signaling cost will result in exclusion from the collective good and a payoff of 0. Signaling institutions will still be able to separate types, as bad types still derive less benefit (from a given increase in cooperation) than good types. However, these will be substantially more costly now that exclusion is possible, as it will be harder to satisfy the bad

type's incentive constraint.

In the presence of exclusion, anonymous signaling institutions are less costly than public ones. On the other hand, public signaling institutions can weed out bad types and prevent crowding. The optimal choice of institution depends on the tradeoff between these concerns, i.e., the choice depends on the magnitude of the crowding effect.

Separating equilibrium in a public signaling institution

Again, in a separating equilibrium, only good types pay the cost σ. Since bad types are identified, and since they increase crowding but make no contributions, their inclusion would harm good types' welfare; hence bad types are excluded.[12] Then the incentive compatibility constraints are, for the good type:

$$\sigma \leq \sum_{g=0}^{N-1} P^G(g) w(\tilde{x}_g, g\tilde{x}_g/N, (g+1)/N)$$

where $\tilde{x}_g$ is the (largest) equilibrium contribution level given $g+1$ included good types, defined analogously with x_g. There is no negative term on the right hand side, since not paying results in exclusion and welfare of 0. The IC constraint for the bad type is

$$\sigma \geq \sum_{g=0}^{N-1} P^E(g) w^E(0, g\tilde{x}_g/N, (g+1)/N). \tag{10}$$

As before, these are compatible when $P^E \approx P^G$. However, the lowest signaling cost does not go to 0 as N grows large, as the cost of being excluded remains high. Instead, the above inequality approaches

$$\sigma \geq \int_0^1 w^E(0, \gamma\bar{x}_\gamma, \gamma) dF(\gamma)$$

[12] Some good types might also be excluded, if their contributions would not compensate for the increased crowding. This issue introduces cumbersome technicalities and is not central to our argument, so we ignore it by fiat: assume that inclusion decisions cannot be probabilistic. Then either all those who pay the cost are included, or none are. (Excluding all good types never increases welfare, since welfare is always non-negative in equilibrium.)

where $\bar{x}_\gamma \equiv \tilde{x}_{\gamma N}$. Welfare for the good type under the cheapest possible signaling institution is then

$$\begin{aligned}&\int_0^1 w(\bar{x}_\gamma, \gamma\bar{x}_\gamma, \gamma)dF(\gamma) - \sigma \\ =&\int_0^1 w(\bar{x}_\gamma, \gamma\bar{x}_\gamma, \gamma) - w^E(0, \gamma\bar{x}_\gamma, \gamma)dF(\gamma).\end{aligned}$$

Thus, good types' benefit is limited by the "entry fee" they must pay to prevent bad types from pooling.

Our main result is that in certain cases, anonymous institutions give higher good type welfare than public ones.

Proposition 3. *When exclusion is possible, and when $w(x,X,P)$ is differentiable in P and $w_3(x,X,P) < \varepsilon$ for low enough ε, the lowest cost anonymous signaling institution yields higher welfare for good types than any public signaling institution. The same holds when $P^E \approx P^G$ and $w^E(0, g\tilde{x}_g/N, (g+1)/N) \approx w(\tilde{x}_g, g\tilde{x}_g/N, (g+1)/N)$.*

Proof. See the Appendix. □

Again, the intuition is simple. The benefit of exclusion comes from reduced crowding. The cost is from having to make signaling more expensive. But any non-zero cost of crowding makes it rational to exclude known bad types, and requires the large extra signaling cost to prevent bad types pooling. So when crowding is not very important, it is better for signaling to be anonymous. Similarly, when bad types want the collective good almost as much as good types, the cost of public signaling cancels out all the gains from the public good, while anonymous signaling remains fairly cheap.

4.5 A minigame

While using a simple "burning money" form for the signaling institution allows for straightforward welfare comparisons, the real world often presents more complex institutions. It

seems natural that the signaling institution might take a similar form to the basic game, since, this should make it easier to separate the types (as good types benefit more their own contribution, it becomes "cheaper" for them to signal). While such institutions do not so obviously destroy resources, they still may be costly. In a real-world economy, there is not a limitless supply of mutually beneficial collective-action projects, and there will ultimately be diminishing returns. Diverting resources from a big project to a smaller one that is more conducive to signaling may lead to inefficiency. We model this form of institution to derive experimental predictions on behaviour; we leave the welfare analysis of such schemes for future work.

Suppose now that the signaling institution takes the same form as the main game, but with payoffs multiplied by a constant $\alpha > 0$. (We call this a "minigame".) We first examine the anonymous case, in which only total contributions are revealed, and seek an equilibrium in which good types play x^* and bad types play 0 in the minigame; thus the number of good types is given by $g = \sum x_i / x^*$.

Good types are playing optimally in the first round and are included in the second round. They could increase second round contributions by upping their first round contributions by some multiple "M" (e.g., giving twice or thrice x^*) to convince others that there are more good types. (Lowering first round contributions will not be appealing, given reasonable out of equilibrium beliefs.) Hence, for an optimal x^*, we require that no such deviation be profitable, i.e.,

$$\sum_{g=0}^{N-1} P^G(g) \left\{ \alpha w(x^*, gx^*/N, 1) + w(x_g, gx_g/N, 1) \right\} \geq \sum_{g=0}^{N-1} P^G(g) \left\{ \alpha w(Mx^*, gx^*, 1) + w(b(gx_{g+M}), gx_{g+M}/N, 1) \right\} \tag{11}$$

for all $M \in \{2, 3, ...\}$ such that $Mx^* \leq 1$. This will hold if N is large enough that the individual effect of donations is negligible, or trivially if $x^* > 1/2$. We assume it holds.[13] (We also assume that the out-of-equilibrium belief after observing $\sum x_i \in (gx^*, (g+1)x^*)$ is that there are only g good types.)

[13] If this were not the case, we could relax the good type condition by examining an equilibrium in which e.g. $x = 1$ is played in the minigame, with similar results.

The incentive compatibility (IC) condition for bad types is

$$\sum_{g=0}^{N-1} P^E(g)\left[\alpha w^E(0,gx^*/N,1)+w^e(0,gx_g/N,1)\right] \geq \sum_{g=0}^{N-1} P^E(g)\left[\alpha w^E(x^*,gx^*/N,1)+w^E(0,gx_{g+1}/N,1)\right]$$

or, rearranging,

$$\alpha \geq \frac{\sum_{g=0}^{N-1} P^E(g)\left[w^E(0,gx_{g+1}/N,1)-w^E(0,gx_g/N,1)\right]}{\sum_{g=0}^{N-1} P^E(g)\left[w^E(0,gx^*/N,1)-w^E(x^*,gx^*/N,1)\right]}. \tag{12}$$

Next, we turn to the public signaling institution case. In a separating equilibrium, bad types will be excluded; again, we assume that good types will never be excluded. We also assume that out-of-equilibrium contributions of $x \neq x^*$ in the minigame result in exclusion. Since individual behaviour is observed, there is no way a good type can mimic *multiple* good types. By equilibrium behaviour, by playing x^* good types both maximize first-round utility and send the strongest possible signal in the first round (i.e., that they are good types for sure). Hence the good type's IC condition is always satisfied. On the other hand, the bad types have an incentive to play x^*both to avoid exclusion and to induce more second-round contribution. Their IC condition is

$$\sum_{g=0}^{N-1} P^E(g)\alpha w^E(0,gx^*/N,1) \geq \sum_{g=0}^{N-1} P^E(g)\left[\alpha w^E(x^*,gx^*/N,1)+w^E(0,gx_{g+1}/N,(g+1)/N)\right]$$

or rearranging

$$\alpha \geq \frac{\sum_{g=0}^{N-1} P^E(g) w^E(0,gx_{g+1}/N,(g+1)/N)}{\sum_{g=0}^{N-1} P^E(g)\left[w^E(0,gx^*/N,1)-w^E(x^*,gx^*/N,1)\right]}. \tag{13}$$

Inspection reveals that (13) is tighter than (12). In other words, just as with a "burning money"-style institution, the lowest α that will sustain a separating equilibrium is higher when the signaling institution is public.

Pooling equilibria with exclusion

To derive predictions for our experiments we examine pooling equilibria in the minigame. Suppose in particular that all types play x_{N-1} in the minigame. Recall that this is optimal for the good types when there are $N-1$ other good types, i.e. when everyone is good; hence it is also maximizes welfare in the minigame when all types contribute. Suppose also that contributing$x < x_{N-1}$ results in exclusion. (Even if $x > x_{N-1}$ does not result in exclusion, neither type will prefer to play it if both are included when playing x_{N-1}, assuming as before that (11) is satisfied.) No players are excluded and in the main game good types play x^* since they have no information about the type distribution. Good types are clearly playing optimally whatever the value of α. The condition for bad types to prefer to pool is

$$\alpha \leq \frac{\sum_{g=0}^{N-1} P^E(g) w^E(0, gx^*/N, 1)}{w^E(0,(N-1)x_{N-1}/N,1) - w^E(x_{N-1},(N-1)x_{N-1}/N,1)} \tag{14}$$

In general the right hand side of(14) will be higher than that of (12). For instance, as N grows large, the right hand side of (12) goes to 0.

Lemma 4. *If α satisfies* (12) *and* (14)*, then a pooling equilibrium is possible in a public signaling institution, while a separating equilibrium is possible in an anonymous signaling institution. If play conforms to these equilibria , then*

(1) The correlation between a player's contributions in the minigame and in the main game is higher under anonymity than with a public minigame;

(2) The probability of exclusion is decreasing in contributions in the public minigame, but is always 0 in the anonymous minigame;

(3) Agents' contributions in the main game are perfectly predictable from contributions in the anonymous minigame, while contributions in the public minigame are not informative at all;

(4) Under anonymity agents' posterior beliefs about other agents are certain, and are perfectly correlated with actions in the minigame. In the public signaling institution, poste-

riors are identical with priors and are uncorrelated with actions in the minigame (along the equilibrium path).

(5) Under anonymity, good types' contributions in the main game are positively correlated with the total level of contributions in the minigame.

(6) When w is concave, $z(\cdot)$ is weakly concave and $w_1(x,X)$ is weakly concave in X, average contributions in the main game are higher after the anonymous minigame.

Proof. Results 1-5 follow from the definitions of equilibrium behaviour. In particular, (4) describes beliefs along the equilibrium path. Off-equilibrium we require that the belief after observing $x \in (0,x^*)$ is low enough to justify exclusion. (5) follows since x_g is increasing in g and g is given by the number of players contributing in the anonymous minigame. In the revealed minigame, technically, total contributions only take on one value of x^*, so we cannot speak of a correlation. If there were some error, contributions would be uncorrelated across the two games.

(6) is proved in the Appendix.

□

Our experimental setup is not identical to our theoretical model, and we do not believe there are only two types of subjets, or that our subjects instantly play signaling equilibria. So we use the predictions of Lemma 4 to guide our hypotheses, rather than literally fitting the model.

5 Experiment

5.1 Design and implementation

The model of the previous section predicts that small symbolic actions (in the "signaling institution") may be informative of players' true preferences and thus of their play in a more important collective action problem (the "basic game"). However, when exclusion is possible,

there is a strong incentive to exclude antisocial types, and therefore unless the signaling institution is rather costly, it may fail to induce separation and thus fail to be informative about players' preferences. As an equilibrium exists, these results could presumably be "manufactured" in the lab through induced payoffs whose distributions are common knowledge and ritual/exclusion institutions designed to guarantee a separating equilibrium only in the anonymous case. This would essentially test subjects' ability to learn and coordinate on the efficient equilibrium.[14] We prefer instead to test our model and its predictions in a more ambiguous environment. In particular, existing work on cooperation claims that different people react differently to material incentives: some are reciprocators while others are selfish. We believe this is a key case in which anonymous signaling institutions may develop. However, if humans do not all have materially self-interested preferences, then they may also fail to conform to other standard economic predictions; this would challenge our game-theoretic explanation of the institutions mentioned in Section 3. Thus, we test our theory using home-grown values (Harrison 2002), by setting up an environment in which a signaling institution of a fixed size is provided, and observing whether players learn to make use of it. If they can do so in the during a brief laboratory experiment, this will increase our confidence that they might also do so in the real world over long time periods, thus giving credence to our explanation. Similarly, if signaling institutions without anonymity fail in the way we expect, it seems likely that real-world institutions may face problems of pandering.

Also, rather than using a pure "burning money" signaling institution, we use a "minigame" as in Subsection 4.5. We expected this to be more intuitive for subjects, and to be better at separating types, since good types have preferences for higher contributions in the minigame even aside from its signaling effect.

To test our theory, we measure whether subjects can use first-round play to detect whether others in their group will cooperate. We compare their inferences of second-round behavior to the actual relationship between a player's first and second-round contribution in both of

[14] If the subjects were not learning optimally, or were coordinating on a less efficient equilibrium, we could limit our subjects to graduate students in economics and force them to read our working paper before playing.

our treatments. Finally, we measure how this interaction evolves over time, and compare anonymous and revealed institutions' success at sustaining cooperation. This question is theoretically ambiguous: "better" signaling will also be better at signaling that there are *not* enough sociable types. This issue is further complicated by the possibility that conditionally cooperative preferences may themselves depend on the outcome of previous rounds (e.g., a sociable type may become antisocial if she was betrayed before).

Figure 2: Experimental design

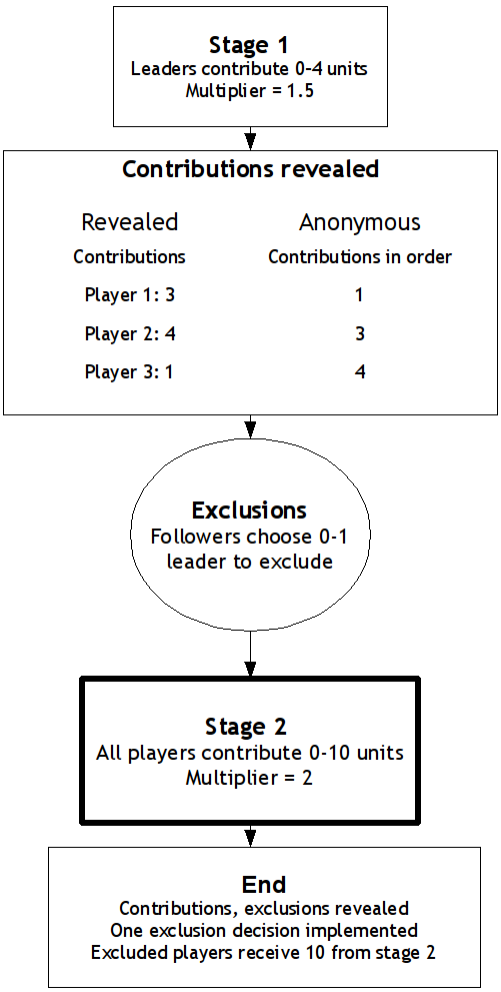


Figure 2 shows our setup. Thirty subjects enter; fifteen are randomly assigned to the *Anonymous*, and fifteen to the *Revealed* treatment. Subjects play 15 repetitions (six repetitions in the pilot version) of a two-stage linear public goods game with exclusion. Subjects

always remain within the same anonymity treatment, i.e., this is a between-subjects design. For each repetition, subjects are randomly assigned to groups of five – we use a "stranger matching" design. Players in each group are randomly numbered from 1 to 5. [15]

The game has two contribution stages, with revelation and an exclusion decision in between these.

Stage 1: The *signaling Institution*

In Stage 1 of this game three subjects (we will refer to these as *leaders*, although we did not use this terminology in the experiment) play a small-stakes public goods game among themselves. Players 1-3 each receive an endowment of 4 points. They choose to contribute between 0 and 4 points to a common fund. Contributed points are multiplied by 1.5 and shared equally among players 1-3. For example, if players 1-3 contribute 3, 1, and 5 respectively, then each player receives (3+1+5)*1.5/3=4.5 points in addition to what they kept.

The marginal per-capita return (mpcr) is $\frac{1}{2}$; thus, the unique equilibrium in dominant strategies with material payoffs is for no one to contribute.

Revelation of Stage 1 Choices

Next, the individual Stage 1 contributions of players 1-3 are shown to all five players. In the *Revealed* condition, these contributions are labelled by player numbers. Continuing the example, all players will see "Player 1: 3 points; player 2: 1 point; player 3: 5 points." In the *Anonymous* condition, contributions are not labelled. In our example, all players will see "Contributions (lowest to highest): 1, 3 and 5 points".

[15] In even repetitions, first players 1-2 are selected for each group, from a pool made up of the previous repetition's players 4-5; then players 3-5 are selected for each group from the remaining players. This ensures a reasonable balance of leader/follower roles across subjects.

Belief elicitation

In repetitions 3, 7, 11, and 15, after stage 1 choices are revealed, we ask each subject to predict what each other member of the group will contribute in Stage 2. Points are given based on the accuracy of their guesses, according to a quadratic scoring rule.[16]

Exclusion decision

Each of players 4-5 ("followers") now chooses who, if anyone, to exclude from the (outcome of the) stage 2 public goods game. She may choose either not exclude any player, or may exclude either player 1, 2 or 3. The exclusion choices and outcome are not revealed until the end of the repetition.[17]

Stage 2: The *Basic Game*

In Stage 2 all five group members receive an endowment of 10 points. We make this endowment larger, like the "basic game" in our model, so as to make the signaling effect of round 1 contributions more important. Each player chooses to contribute between 0 and 10 points to a common fund.

End-of-Game Revelation

One of the two followers' exclusion decisions is chosen at random. If the decision was to exclude a player, this leader is excluded: he receives 10 points from round 2 and his round 2 contributions to the common fund are not counted. If the decision was not to exclude any player, all five players' contributions are counted. Points contributed by non-excluded players

[16] At the end, we choose a single guess from a single subject from a single repetition for this payment. In addition to the other payments this subject receives 20 euros less one fifth the square of the error in this prediction.

[17] This allows us to gather data on excluded subjects' round 2 choices. It also minimizes the effect that learning of an exclusion might have on other subjects' decisions. Although this effect may not disappear completely, as the perceived probability of an exclusion may vary, such a bias would be likely to work *against* our finding of higher contributions under anonymity. If exclusions are more likely in the revealed treatment, and subjects contribute more after an exclusion (e.g., because the mpcr is slightly higher and the "worst player" is gone) then contributions in the revealed treatment are biased *downward.*

are doubled[18] and shared equally among these non-excluded players.

Points earned by each player in stage 1 and stage 2 are recorded. At the end of the experiment repetitions are randomly selected for actual payments; to avoid wealth effects we independently draw one repetition for stage 1 payments and one repetition for stage 2 payments, paid at a rate of 1 Euro = 1 point.. Each player is shown:

(1) The contributions of all players in round 2, labeled by player number.

(2) his points earned in Stages 1 and 2 (with the excluded player learning that he earned 10 points in Stage 2), and his total points earned;

(3) the contributions of all players in round 1, labeled by player number in the Revealed treatment, unlabeled and ordered in the Anonymous treatment;

(4) the two exclusion decisions made by players 4-5 and

(5) the player chosen to be excluded from round 2.

Implementation details

The experiment was run at the Max Plank Institute in Jena, Germany. The subjects came from the usual convenience sample, mainly university students, recruited by standard procedures. We ran five sessions with 30 subjects per session, 15 in each treatment. In addition to the earnings mentioned, participants received a showup fee of EUR 2.50.

5.2 Predictions

Let C_1 be the total contributions made in round 1 and C_2 be those made in round 2. Let c_j^i be the contribution made by player $i \in \{1,2,3,4,5\}$ in round j, so that $C_1 = c_1^1 + c_1^2 + c_1^3$ and $C_2 = \sum_{i=1}^5 c_2^i$, and C_j^{-i} represents the sum of contributions made in round j by players other than i. Let $\{C_1\} = \{c_1^1, c_1^2, c_1^3\}$, $\{C_2\} = \{c_2^1, c_2^2, ..., c_2^5\}$ and $\{C_j^{-i}\}$ the set of all contributions in round j other than i's contribution, and $\{E^i C_j^{-i}\}$ the vector of i's expectations of these. Let

[18]This return rate is chosen so that, if one player is excluded, the MPCR is $\frac{1}{2}$, as in the first round. If no one is excluded the MPCR is $\frac{2}{5}$, only slightly lower. Thus, players' preferences for contribution in each round will plausibly be quite similar, ignoring exclusion-risk and signaling effects.

e^i be an indicator that player $i \in \{1,2,3\}$ is excluded, and e^0 an indicator that *some* player is excluded. Let T represent an indicator that takes the value $T = 1$ in the identified treatment and $T = 0$ in the anonymous treatment.

For empirical purposes, we expect a range of types, rather than the model's two types, with more cooperative players willing to contribute more for a given level of others' contributions. Nevertheless, in equilibrium, some or all types may pool in the first round. We maintain as an assumption that our chosen payoff sizes cannot sustain full separation when contributions are revealed, but can separate types to alarger extent when they are anonymous.[19] Therefore, under anonymity, contributions in round 1 will be informative about type, and hence predictive of contributions in round 2. Under the revealed treatment, the correlation will be weaker or non-existent, depending on whether there is full or partial pooling. We do not expect a negative correlation in either case. Thus, our empirical prediction is a weakened version of Lemma 4 (1):[20]

Prediction 1. $\rho(c_1^i, c_2^i | T = 0) > \rho(c_1^i, c_2^i | T = 1) \geq 0$.

Turning to the exclusion decision, since all types of players have a material incentive to exclude those who will contribute less than the average, players will exclude less the higher the round 1 contribution.[21] Specifically, players will tend to choose to exclude the player who contributed least, when he can be identified.

Prediction 2. *$E(e^0|\{C_1\})$ is non-increasing in the elements of $\{C_1\}$, and is strictly decreasing in* $\min\{C_1\}$.

Since exclusion can be targeted in the revealed treatment, we expect that lowering one's contribution increases the risk of one's own exclusion *more* in this case (a weakening of

[19] We chose the payoff sizes by intuitive judgment rather than by theory.

[20] We expect a non-negative relationship because if selfish types contributed more in round 1, then by lowering their contributions they could both increase their payoff and look more like reciprocators, hence reducing their probability of exclusion. If there were full pooling, round 1 donations would be completely uninformative (although out of equilibrium beliefs would still be monotonic in round 1 donations under a reasonable equilibrium restriction). Realistically we do not expect a pooling equilibrium to arise immediately, nor do we expect it to be played all the time.

[21] If there is full pooling in round 1, this will be driven by out-of-equilibrium beliefs.

Lemma 4 (2).

Prediction 3. $\partial E(e^i|c_1^i, T=1)/\partial c_1^i < \partial E(e^i|c_1^i, T=0)/\partial c_1^i \leq 0$, *the probability of a player's exclusion is more affected by that player's contribution in the revealed treatment than in the anonymous treatment (but both effects are non-positive).*

This in turn suggests there will be a greater incentive to pool in the revealed treatment, justifying Prediction 1These differences should also be reflected in players' subjective expectations $E^j(\bullet)$. Because it is more likely to separate types, in the anonymous treatment expectations of round 2 contributions should respond more to round 1 contributions, and expectations should be more accurate (relative to the revealed treatment). Hence, motivated by Lemma 4.3 and 4.4, we make two more predictions:

Prediction 4. $0 \leq \partial(E^j(c_2^i|c_1^i, T=1)/\partial c_1^i < \partial(E^j(c_2^i|c_1^i, T=0)/\partial c_1^i$ *for* $i \in \{1,2,3\}$ *and* $j \in \{1,2,3,4,5\}$.

Prediction 5. $\rho(c_2^i, E^j(c_2^i|c_1^i)|T=1) < \rho(c_2^i, E^j(c_2^i|c_1^i)|T=0)$ *for* $i \in \{1,2,3\}$ *and* $j \in \{1,2,3,4,5\}$.

Next we turn to the effect on other players' round 2 contributions. Following assumption 3, a conditionally cooperative player's optimal round 2 contribution is never decreasing in her expectation of others' contributions, and we will assume that it is increasing over some range. Hence, average contributions in round 2 increase in the expectation of others' contributions: this yields Prediction 6.

Prediction 6. $E(c_2^i|\{E^iC_2^{-i}\}, T=0)$ *is nondecreasing in – and increasing over some range of – the elements of* $\{E^iC_2^{-i}\}$.

By Prediction 4 the expectation of a (leader) player's round 2 contribution is increasing in her round 1 contribution in the anonymous case; combining this with Prediction 6 yields Prediction 7 (similar to Lemma 4.5).

Prediction 7. $E(c_2^i|\{C_1^{-i}\}, T = 0)$ *is nondecreasing in – and increasing over some range of – the elements of* $\{C_1^{-i}\}$.

Finally, reflecting Lemma 4 (6), we predict that the extra information in the Anonymous treatment will, on average, allow for greater contributions by reducing uncertainty. Since VCM experiments (without punishment) have consistently found contributions decline over repetitions, we suspect that the equilibrium *without* a succesful signaling institution will have low cooperation levels; hence signaling should be an improvement.[22]

Prediction 8. $E(C_2|T = 0) > E(C_2|T = 1)$.

5.3 Results

Overview

We first summarize the overall giving patterns by repetition and treatment. While Stage 1 investments remained fairly constant across repetitions (at about 50% of the Stage 1 endowment) in both treatments, stage-two investments began higher and declined much less over the 15 repetitions under the *Anonymous* treatment, remaining at above 30% on average while falling to around 10% in the *Revealed* treatment. We argue that the anonymous first stage made this persistent cooperation – which is unusual in the absence of punishment (Ledyard 1993) – possible. Figure 4 shows that these patterns are similar across sessions. Stage 2 average investment in the final stage is strictly lower in the revealed case for *all* sessions. Stage 1 investment is more mixed, but mainly lower in anonymous treatments.

[22]This accords with the interpretation of Holt and Laury (n.d.), who note that, "if cooperative gain-seekers [Brandt and Schram, 1996] systematically overestimate the number of others of this type in their group" this could explain the decline in cooperation over time in VCM contributions. Lemma 4 gives conditions for an anonymous signaling institution to increase contributions.

Figure 3: Mean Contributions by Repetition by Treatment

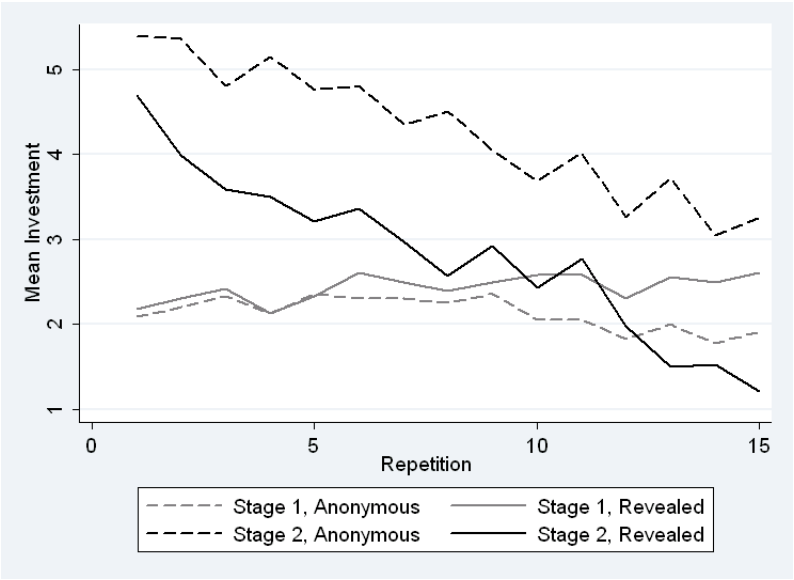
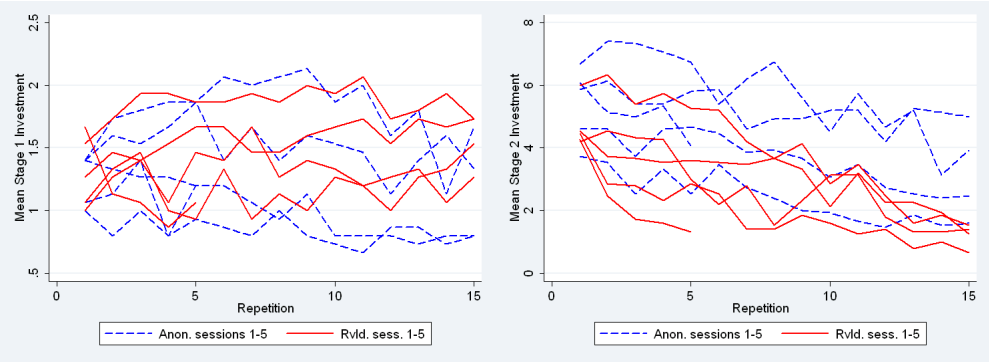
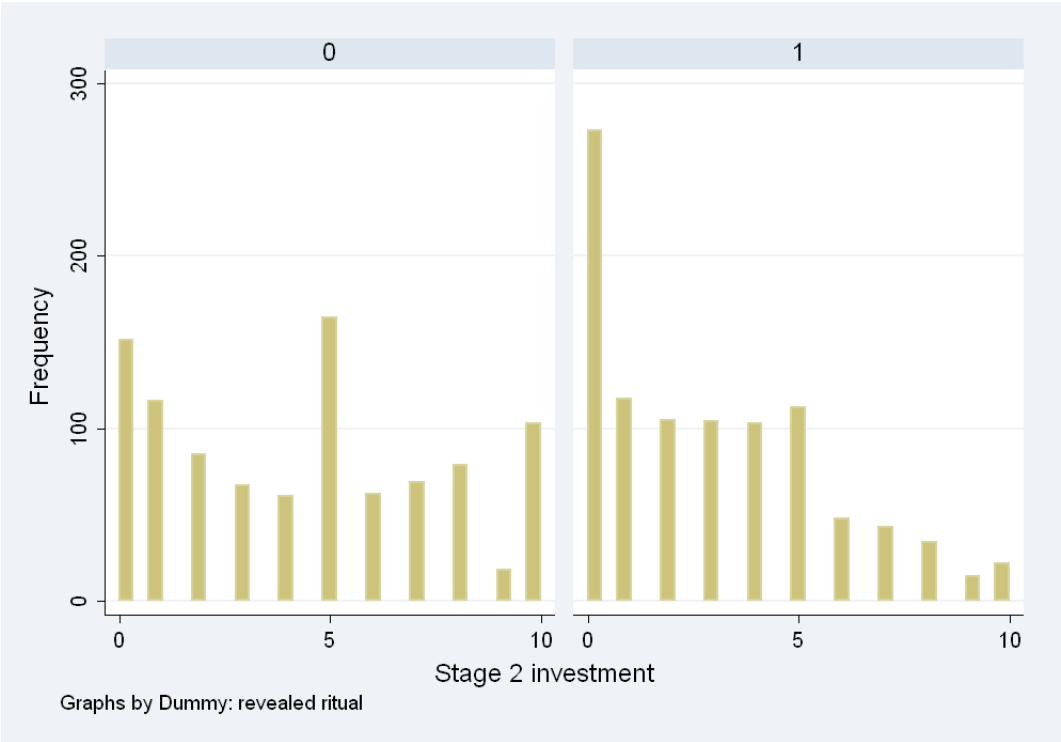


Figure 4: Sessions: Mean Contributions by Repetition by Treatment



The difference between Stage 2 investments was statistically significant in both parametric and nonparametric tests (Wilcoxon rank-sum $p = 0.07$, taking the treatment/session as the unit of observation). This result supports Prediction 8. Figure 5 shows frequencies of different Stage 2 investments by treatment. The spike at contributions of 5 in the anonymous condition suggests that participants may be coordinating successfully on a "reasonable" focal point. As we shall see, the data support this interpretation.

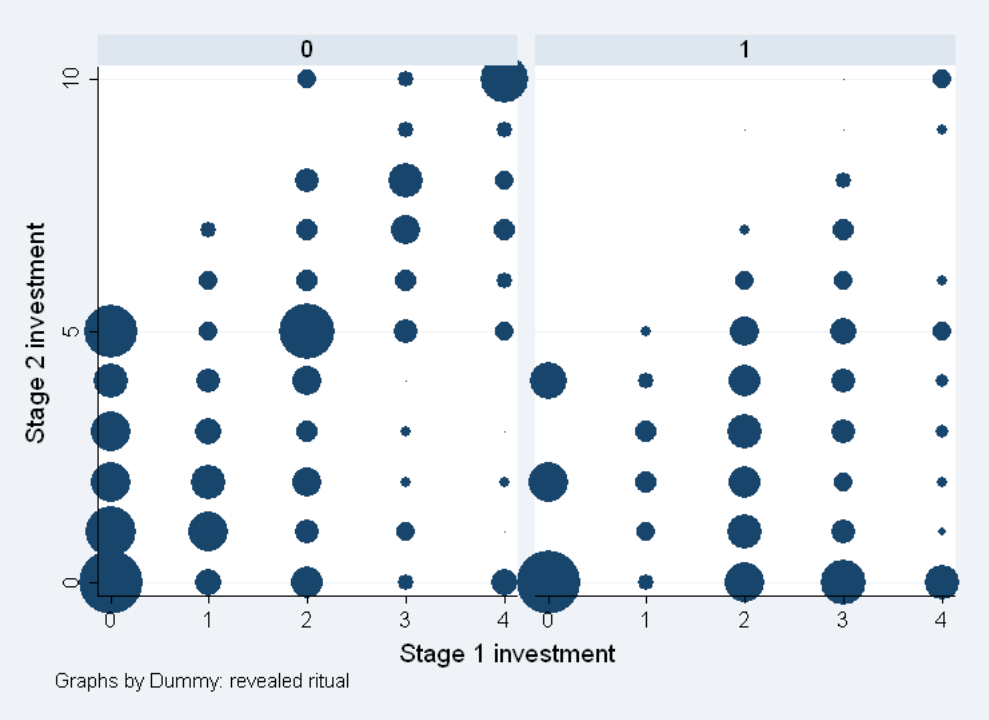
Figure 5: Histogram: Stage 2 Contributions by Treatment, all Repetitions



Round 1 and round 2 behaviour

Figure 6 shows frequencies of Stage 2 contributions by Stage 1 contributions for each treatment (bubble width indicates number of observations), for the later repetitions, presumably after some strategic learning has taken place.

Figure 6: Stage 1 investment by stage 2 investment (repetitions 8-15). Left=Anonymous, right=Revealed



Prediction 1 appears to hold: there is a positive correlation between giving in the stages, but the correlation is much stronger in the anonymous treatment. In particular the revealed treatment shows a lot of high Stage 1 contributions followed by low Stage 2 contributions, which suggests attempts to avoid exclusion. Next we decompose the variance into its explained and unexplained components, reporting marginal[23] and total sums of squares.

In the revealed treatment a subject's stage 1 investment has little relation to her stage 2 investment. In contrast, in the anonymous treatment first-stage investment is highly informative. It is not just the *presence* of an anonymous stage 1 contribution that matters, but also its magnitude; a 1 ecu investment explains little, while larger investments matter a great deal. Ultimately, the vast majority of the explanatory power of first stage investment is via individual heterogeneity. In the final two columns, after conditioning on subject-specific effects, first stage investment explains little of the remaining variation for either treatment. While the

[23]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 TSS only if the variables are exactly orthogonal.

Table 1: Analysis of variance of Stage 2 contributions by Stage 1 contributions

Partial (marginal) sums of squares:	Anon.	Revealed	Anon. later	Rvld. later	Anon.	Rvld.
1 ecu invt.	14	14	35	7	16	21
2 ecu invt.	211	20	199	9.1	37	24
3 ecu invt.	605	42	428	15	36	28
4 ecu invt.	1008	88	497	19	75	44
Subject effects					2038	1560
Model Degrees of freedom	4	4	4	4	78	78
Observations	585	585	288	288	585	585
Model SS	1937	133	849	27	3975	1694
Total SS	6172	3997	2846	1519	6172	3997
R-sq.	.31	.033	.3	.018	.64	.42

'Later' refers to stages 8-15

subject effects explain roughly one third of the variance in contributions in both treatments, these effects are not observable to the subjects, only to the econometrician (ex-post); hence the informativeness of first-stage investment matters. [24]

Econometric Discussion

Above, we reported some simple results at the level of the treatment/session; clearly such observations are strictly independent, but do not fully exploit the information in the data. Our design uses "imperfect" stranger matching; subjects play many repetitions in a limited pool. As in all such experiments the observations (subjects/repetitions) are not completely independent, and play may be affected by experience in earlier repetitions. We deal with this potential dependence in several ways. Our regressions estimate robust standard errors, clustered at either the subject level or the treatment/session level, as appropriate to the specification.[25] Where noted, we also use a set of four control variables for a subject's (i's) experience: (i) C_1^{-i} , (the average stage one investment of subjects in i's group other than i)

[24] Our (Poisson) regression analysis of stage 2 contributions (available by request) yielded comparable results: the coefficient on Stage 1 investment is significant and positive in the anonymous case, but significantly smaller (and sometimes insignificant) in the revealed case. Again, a 1 ecu investment explains little, while the coefficient on the larger anonymous investment dummies are significant and suggest a nonlinear relationship.

[25] We do not estimate random-effects models because these rely on additional assumptions on the error structure, such as homoskedasticity of the individual error terms. In general, the results hold with or without random effects estimation.

for i's the previous repetition, (ii) the same for C_2^{-i}, and (iii,iv) the means of C_1^{-i} and C_2^{-i}, respectively, over *all* of i's the previous repetitions. [26]

Exclusion

Exclusion occured in both treatments, but was significantly more common in the Revealed treatment, particularly in earlier repetitions.

Figure 7: Exclusion Rates by Repetition by Treatment

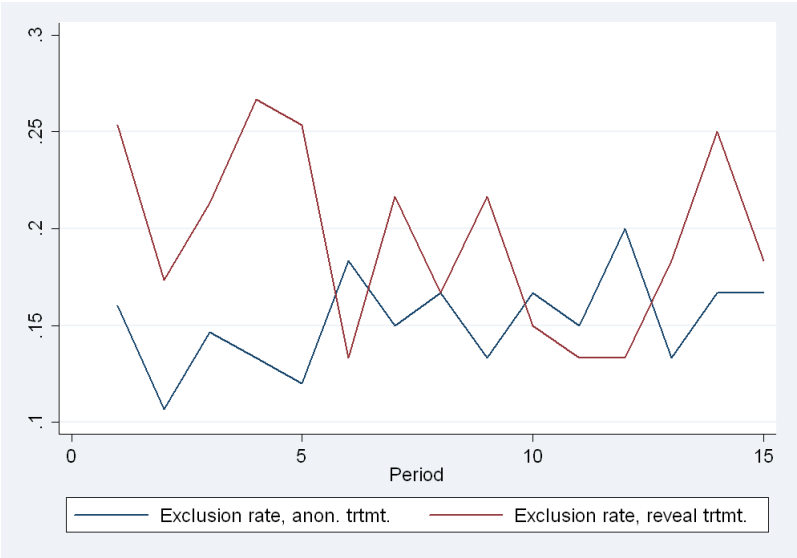


Table 2 gives Probit regressions for a follower's decision whether or not to exclude one of the leaders. As Prediction 2 implies, the probability of an exclusion decreases in the minimum amount contributed by a leader; in general, this holds for both treatments. For *early* repetitions, the minimum gift has a negative and sometimes significant impact on the probability of an exclusion for both treatments. However, for the revealed treatment the effect is significantly greater (again following Prediction 2) and persists even through the later repetitions.[27]

[26] This specification was chosen for parsimony and based on some preliminary tests. The simple lag term is motivated by the idea that memory has a recency bias. Regressions allowing an intercept for sessions/treatments yielded similar results.

[27] In some regressions anonymous median gifts have a negative and significant coefficient, while the coefficient on revealed median giftsis positive and significant; while we speculate that this is because greater contrast between the contributions makes the exclusion decision harder in the former case and easier in the latter, we made no theoretical predictions for these coefficients.

Table 2: Probit regressions: decision to exclude someone

Dependent variable = Dummy: subject chooses to exclude someone.	(1)		(2)		(3)		(4)	
	All Repetitions		All Repetitions		Repetitions 8-15		Repetitions 8-15	
Minimum St. 1. Invt.	-0.075	(0.049)	-0.13*	(0.056)	-0.012	(0.065)	-0.10	(0.067)
Rvld × Min St. 1 Invt	-0.17*	(0.081)	-0.18*	(0.088)	-0.20+	(0.11)	-0.15	(0.11)
Median St. 1. Invt.	-0.011	(0.039)	-0.079+	(0.041)	-0.058	(0.053)	-0.11*	(0.054)
Rvld × Med. St. 1 Invt.	0.16*	(0.067)	0.19*	(0.080)	0.22*	(0.11)	0.21+	(0.12)
Range St 1. Invts.	-0.011	(0.041)	-0.040	(0.044)	0.028	(0.053)	-0.0027	(0.054)
Rvld × Range St. 1 Invts.	0.062	(0.062)	0.055	(0.071)	0.13	(0.092)	0.13	(0.093)
Revealed Trtmt.	-0.054	(0.17)	-0.018	(0.19)	-0.24	(0.22)	-0.072	(0.24)
History & Lag 1 Var's	No		Yes		No		Yes	
Observations	780		660		384		384	

+ p<0.10, * p<0.05, ** p<0.01
Estimates reported as marginal effects at mean values.
Standard errors (clustered by subject) in parentheses.

To test Prediction 3, we examine the impact of a leader's gift on the probability that she is excluded. Table 3 demonstrates that the probability a leader was excluded[28] varied inversely with her Stage 1 investment, and this effect was much stronger in the revealed treatment. More intuitively, the probability of exclusion was 78%, 53%, 12%, 6%, and 4% given revealed Stage 1 contributions of 0,1,2,3, and 4 respectively, with an even steeper slope in later stages (but fewer small gifts, as the leaders learn this makes them an easy target for exclusion). In other words, in the revealed treatment investing more than 2 in stage 1 offers little further protection against exclusion, hence it may be seen as a signal of a "very" good type.

Beliefs

Table 4 regresses players' predictions about leader players' round two contributions on the leaders' actual round one contribution. The coefficient of first-stage investment is somewhat lower (significantly in columns 1 and 6) in the revealed case, although the summed coefficient remains significant. This supports Prediction 4. In line with Prediction 5, subjects'

[28]In the anonymous case, 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.

Table 3: Poisson regressions: exclusion votes against a player

Dependent variable = Number of votes [*] to exclude subject (in single repetition).	(1)		(2)		(3)	
	All Repetitions		Repetitions 8-15		All Repetitions	
St. 1 Investment	-0.13**	(0.049)	-0.17*	(0.071)		
Rvld × Invt.	-0.80**	(0.13)	-0.72**	(0.20)		
Repetition	0.015	(0.012)	-0.0047	(0.025)	0.014	(0.012)
Rvld × Reptn.	-0.012	(0.020)	0.035	(0.036)	-0.010	(0.022)
Invested 1 ecu					-0.24*	(0.12)
Rvld × Invt. 1 ecu					-0.29	(0.18)
Invested 2 ecu's					-0.32+	(0.17)
Rvld × Invt. 2 ecu's					-1.45**	(0.44)
Invested 3 ecu's					-0.56**	(0.20)
Rvld × Invt. 3 ecu's					-2.28**	(0.57)
Invested 4 ecu's					-0.52**	(0.17)
Rvld × Invt. 4 ecu's					-2.73**	(0.24)
Constant	-1.09**	(0.12)	-0.64+	(0.39)	-1.01**	(0.15)
Session/Trtmt. Dummies	Yes		Yes		Yes	
Observations	1170		576		1170	

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

[*] Conditional expectation simulated for anonymous case; see footnote.

Poisson coefficients: marginal effects reported.

Robust (clustered by session/treatment) standard errors in parentheses.

guesses were significantly better in anonymous treatments than in revealed treatments, with correlations to targets' choices of 0.39 and 0.26 respectively, and predictions did not strongly improve in later stages (details by request). Subjects were overoptimistic in both treatments, significantly so only in the revealed treatment, in which they overpredicted by an average of 0.85 ecus.

Stage 1's effects on others' stage 2 investments

By prediction 6, we expect (in both treatments) a player's stage 2 investments to increase in her expectation of others' investments. However, the prediction itself may be correlated to subject-specific unobservables (e.g., more generous people may be more optimistic about others' generosity); hence we control for a subject-specific effect.[29]

[29] 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 are not sensitive to this.

Table 4: Poisson regressions: predictions for leaders

Dependent variable = Prediction of target's stage 2 investment						
	(1)	(2)	(3)	(4)	(5)	(6)
	All reps, Poisson	...	Reps 11,15	...	Rep 15	...
Target St. 1 Inv.	0.34**	0.25**	0.36**	0.24**	0.37**	0.27**
	(13.61)	(9.34)	(10.61)	(5.74)	(8.63)	(5.65)
Tgt. St. 1 Inv × Rvld.	-0.069+	-0.038	-0.091	-0.017	-0.22*	-0.15
	(-1.69)	(-0.97)	(-1.21)	(-0.21)	(-2.35)	(-1.55)
Repetition	-0.0077	0.0071				
	(-1.26)	(1.07)				
Rvld. × Repetition	-0.050**	-0.041**				
	(-4.65)	(-3.76)				
Dummy: revealed ritual	0.34*	0.33*	-0.25	-0.14	-0.044	0.0080
	(2.42)	(2.39)	(-1.12)	(-0.61)	(-0.17)	(0.03)
Constant	0.70**	0.19	0.55**	0.14	0.50**	0.079
	(7.87)	(1.58)	(5.63)	(1.23)	(4.21)	(0.48)
History & Lag 1 Var's	No	Yes	No	Yes	No	Yes
Observations	1152	1152	576	576	288	288
Sum: invt. & rvld.	.27**	.21**	.27**	.20**	.15+	.13+

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

Session 1 excluded due to error in prediction instructions.

Poisson coefficients: marginal effects reported.

Robust std. err. (clustered by subject) in parentheses.

In anonymous treatments predictions were for (e.g.,) "the guy who contributed 4 ecus."

Table 5: Determinants of followers' stage 2 investment: Measuring Conditional Cooperation

	(1)	(2)	(3)	(4)
	All reps	All reps	Reps 8-15	All reps
	Poisson regressions of follower subjects' investments			
Min guess (tgt: leader)	0.11*			0.054
	(0.056)			(0.052)
... × Rvld. Trtmt.	0.065			0.080
	(0.087)			(0.062)
Med. guess (target:leader)	0.021			-0.029
	(0.068)			(0.060)
Max. guess (tgt:leader)	0.084			0.155**
	(0.061)			(0.044)
Guess (tgt:follower)	0.053			0.075**
	(0.038)			(0.022)
Num. Ldrs. Inv. 2+ [a]		0.205*	0.257+	0.104
		(0.092)	(0.139)	(0.180)
... × Rvld. Trtmt.		-0.081	-0.736*	-0.187
		(0.161)	(0.333)	(0.673)
Num. Ldrs. Inv. 3+ [b]		0.125+	0.149	-0.187+
		(0.066)	(0.103)	(0.112)
... × Rvld. Trtmt.		0.087	0.020	0.153
		(0.105)	(0.167)	(0.162)
Repetition	0.026	-0.035**	-0.036	-0.006
	(0.024)	(0.010)	(0.027)	(0.024)
Rvld × Reptn.	-0.030	-0.046**	-0.115**	0.013
	(0.029)	(0.015)	(0.038)	(0.025)
Final Rep.	-0.255	-0.115	-0.023	-0.157
	(0.210)	(0.170)	(0.171)	(0.246)
Dummy: Rvld. Trtmt.		-0.037	2.141**	-0.347
		(0.336)	(0.765)	(1.303)
Constant		1.184**	1.079*	0.119
		(0.196)	(0.492)	(0.392)
Subject-Fixed Effects	Yes	No	No	No [c]
Observations	125	780	384	192
P-val: F test, guesses	0.020			0.000
Sum coef: Num. 3+ Ldrs., Anon		0.331**	0.407*	-0.083
Sum coef: Num. 3+ Ldrs. Rvld		0.338*	-0.309	-0.116

Standard errors in parentheses

"Target" indicates the target of subject's prediction of round 2 gift for that repetition (3,7,11, or 13).

Pilot session dropped from regressions with subjects' prediction variables.

[a] Count of leaders (in group/rep.) who invested 2 or more in st. 1; excludes min. invt.

[b] ... 3 or more

[c] FE yields same qualitative results but little conditional variation 'in Num Ldrs', hence wide std. errors

The first column of table 5 measures the relationship between a player's stage 2 investment and her guesses of others' contributions, a direct measure of conditional cooperation. We limit this analysis to follower players only to rule out a direct reciprocity motive; leaders might feel generous or bitter to other leaders depending on the first-stage play, and their play in the second stage may also be influenced by their perceived probability of being excluded. We allow the slope in the minimum guess to vary by treatment; minimum givers in the *revealed* treatment are likely to be excluded, and an exclusion is more likely the lower the gift. As usual, the regressions in this table control for a differential trend across repetitions; here we also include a "final repetition" dummy to allow for an end-game effect. This first regression shows clear evidence of conditional cooperation:[30] the coefficients on each of the guesses (the lowest, middle, and highest guesses for leader subjects' round 2 investments, and the guess for the other follower subject) are positive, and these are jointly significant in an F-test at the 5% level.

By prediction 7, in the anonymous treatment a player's stage 2 investments tend to increase in others' stage 1 investments. The probability of exclusion is nonlinear in stage 1 investment; hence subjects' inferences from stage 1 behavior may also be nonlinear. In columns 2 and 3 we consider three categories of first stage investment: 0-1 ecus, 2 ecus, and 3-4 ecus. In light of the observed probabilities of exclusion, we formed these categories to potentially reflect bad types who separate, pooling outcomes, and "very" good types, respectively. We measure the direct effect of the number of leaders (0,1, or 2) in each of these categories on the followers' contributions. Since, as mentioned, the (revealed) leader who gives the least is likely to be excluded we do not include the lowest gift in this count.

Column 2 shows a strong positive effect of the number of leaders donating at least 2 on the followers' contributions, and an even stronger impact of donations of 3 ecus or more.

[30]Our interpretation of these results as conditional cooperation might be critiqued on the ground that the subject may first choose how much to invest and her prediction for may be an ex-post rationalization of this choice (see Fehr and Schmidt (...)). In response we first note that our subjects guesses are financially motivated. We second point to our evidence (below) that followers also respond to stage 1 contributions. Finally instrumental variables regressions (available by request) using others' stage 1 investments as instruments for average predicted stage 2 contributions strongly support our results.

As column 3 shows, these effects persist into the later stags for the anonymous treatment, but dissapear in the revealed case (with a significant differential), becoming negative (but statistically insignificant) in sign in net. This suggests that followers grow less confident in revealed first-stage investments as predictors of leaders' second-stage behavior, and thus cease to respond positively.

These regressions are "reduced form"; our theory predicts that effect of first stage on the followers occurs *only* through followers' expectations of stage 2 contributions. Column 4 tests – and fails to reject – this interpretation by regressing followers' investments on both their guesses and the *actual* leader investment counts.[31] The coefficients on the guesses themselves remain largely positive and significant in net. However, after controlling for guesses, the coefficients on the leaders' investment variables are smaller and are no longer positive and significant (for either treatment).

Discussion

The results presented are consistent with the predictions motivated by our model; the anonymous ritual seems to have served as an effective coordination device for conditional cooperators. Overall contribution levels in Stage 2 were significantly and substantially higher in the anonymous treatment. The data also supports our account of the mechanism behind this: leaders often contributed to avoid exclusion under the revealed treatment; thus contributions were more closely correlated between the stages under anonymity; as a result beliefs were more accurate. In later stages followers' gifts continued to vary positively with *anonymous* leaders' gifts, but this relationship dissapeared in the revealed treatment. This suggests an explanation for the relative decline in cooperation in the latter case: good types could no longer confidently identify other good types, and this increasing uncertainty reduced stage 2

[31] We also ran this same regression controlling for a subject-fixed effect, for the reasons previously mentioned. This yielded the same qualitative result but relied on a very small number of observations, since the fixed effect could only be identified for subjects who were followers in multiple prediction stages and invested a positive amount in each. Hence this data yielded very little conditional variation in the "Num Ldrs" variables, few degrees of freedom, and very wide standard errors for the corresponding coefficients.

investments. The robust positive relationship between subjects' investments and their predictions of others' investments supports our story of conditional cooperation. This laboratory experiment was a demanding setting for our theory, since it required subjects to understand and respond to the strategic incentives to manipulate contributions over a few repetitions in a brief timespan. We are correspondingly more confident that over much longer timescales, players could learn from anonymous institutions.

6 Conclusion

In societies without large-scale markets, and in areas that the market does not reach (e.g., inside the firm itself), others' character and intentions towards us may be vital for our success and survival. In these contexts it is crucial to be able to gauge others' character, so we can choose who to interact with and how much to invest in these interactions. This in turn gives some individuals a strong incentive to conceal their true character. Certain rituals can be seen as institutions which provide a forgiving environment, inducing people to act in accordance with their true character by obscuring their identity behind the screen of the collective. In this paper, we developed a model showing how anonymous rituals can foster greater cooperation than revealed ones, and demonstrated in a laboratory public goods game that anonymous contributions can lead to subsequent higher contributions and thus increase participants' welfare.

Further empirical work is needed to establish whether this logic explains the survival of of *particular* rituals and institutions. Still, we note that cultural forms such as song seem especially well-suited for preserving the anonymity of participants and shirkers. We hope this approach will be of interest to anthropologists and sociologists of religion, as well as practical researchers and policy-makers looking for mechanisms to foster voluntary cooperation.

Appendix

Proof of Lemma 1

Proof. Set $w^i(x,X) = w(z^{-1}(X),X) + i[w(x,X) - w(z^{-1}(X),X)]$ for $i > 0$. (z^{-1} is well defined on $[0,1]$, since $z(\cdot)$ is strictly increasing and since $w_1(0,0) > 0, w_1(1,1) < 0$. In fact, $z^{-1}(X) = b(X)$.) It is easy to show that $w^i(x,z(x)) = w(x,z(x))$, and that $w_1^i(x,X) = iw_1(x,X)$. Therefore, the sets of solutions to(4) and (6), i.e. of equilibria with and without knowledge, do not vary with i, and expected welfare when g is known is also unchanged. That the other Assumptions for w continue to hold for w^i is also straightforward. (To show $w_2^i > 0$, differentiate w^i to get $w_2^i(x,X) = iw_2(x,X) + (1-i)w_2(z^{-1}(X),X) + (1-i)w_1(z^{-1}(X),X)(z^{-1})'(X)$, and observe that the last term is zero by definition of z.)

We wish to show that

$$\sum_{g=0}^{N-1} P^G(g)[w^i(x_g, gx_g/N) - w^i(x^*, gx^*/N)] > 0 \tag{15}$$

for high enough i. First observe that if $x_g \geq x^*$, then $w^i(x_g, gx_g/N) - w^i(x^*, gx^*/N) \geq 0$. For

$$\begin{aligned}
& w^i(x_g, gx_g/N) - w^i(x^*, gx^*/N) \\
= \; & w^i(x_g, gx_g/N) - w^i(b(gx^*/N), gx^*/N) + w^i(b(gx^*/N), gx^*/N) - w^i(x^*, gx^*/N) \\
= \; & \int_{b(gx^*/N)}^{x_g} \left\{ \frac{d}{dx} w^i(x, z(x)) \right\} dx + \left[w^i(b(gx^*/N), gx^*/N) - w^i(x^*, gx^*/N) \right]
\end{aligned}$$

where the integral is taken along $z(x)$ (since $z(b(gx^*/N)) = gx^*/N$). Now the second term in brackets is non-negative by $b(\cdot)$ a best response (and positive whenever $x^* \neq x_g$. On the other hand, since $w_1^i(x,z(x)) = 0$, $\frac{d}{dx} w^i(x,z(x)) = w_2^i(x,z(x))$ and this is positive, so the integral is non-negative. Thus the whole term is non-negative. (This whole step holds for any w^i including $w^1 = w$.)

Next for $x_g < x^*$ we have

$$\begin{aligned} & w^i(x_g, gx_g/N) - w^i(x^*, gx^*/N) \\ &= w(x_g, gx_g/N) - w^i(x^*, gx^*/N) \\ &\text{(since } (x_g, gx_g/N) \text{ lies on the } z(\cdot) \text{ curve)} \\ &= w(x_g, gx_g/N) - w(z^{-1}(gx^*/N), gx^*/N) + i\left[w(z^{-1}(gx^*/N), gx^*/N) - w(x^*, gx^*/N)\right] \end{aligned}$$

and as $i \to \infty$ the sign of this depends on the sign of the term in square brackets, which is positive since $z^{-1} = b$ and $b(gx^*/N)$ is the unique best response to gx^*/N.

Thus for i high enough the terms of (15) are always non-negative, and are positive whenever $x_g \neq x^*$. □

Another sufficient condition for knowledge of g to improve good type welfare

Lemma 5. *Suppose that Assumption 8 holds. If $w_1(x,X)$ is weakly concave in X for every x, $z(\cdot)$ is weakly concave, and $w_{22}(x,X) < \varepsilon$ for all x, X and small enough ε, then* (5) *is strictly higher than* (3).

Proof. To show that these conditions are compatible consider $w(x,X) = \alpha x - \beta x^2 + \gamma xX - \delta X^2$, for $\alpha, \beta, \gamma > 0$. Here $w_{11} = -2\beta, w_{12} = \gamma$ and $w_{22} = -2\delta$. w is strictly concave if $4\beta\delta > \gamma^2$. $w_1(x,X) = \alpha - 2\beta x + \gamma X$ and is linear, thus weakly concave in X. $z(x)$ solves $w_1(x, z(x)) = 0$ and is given by $z(x) = \frac{2\beta x - \alpha}{\gamma}$ which is linear, so weakly concave. For β large enough and δ small enough, our conditions will be satisfied.

1. Write $\bar{g} = \sum_{g=0}^{N-1} P^G(g)g$ for an agent's interim expected number of other good types, given that the agent knows she is good (but before she observes any actions). Now

$$\sum_{g=0}^{N-1} P^G(g) w(x^*, gx^*) < w(x^*, \bar{g}x^*) \tag{16}$$

by strict concavity of w.

On the other hand, by concavity of w_1 in its second argument, we have $w_1(x^*, \bar{g}x^*) \geq \sum_{g=0}^{N-1} P^G(g) w_1(x^*, gx^*) = 0$, for any x^*satisfying the equilibrium conditions. But then we must have $x^* \leq x_{\bar{g}}$, since $(x^*, \bar{g}x^*)$ is within the upper level set where $w_1 \geq 0$ to , and $x_{\bar{g}}$ is the largest value such that $w_1(x_{\bar{g}}, \bar{g}x_{\bar{g}}) = 0$.

Now, we get the following chain: $w(x^*, \bar{g}x^*) \leq w(x^*, \bar{g}x_{\bar{g}}) \leq w(x_{\bar{g}}, \bar{g}x_{\bar{g}})$. The first inequality holds by welfare increasing in total contributions, the second by optimality of $x_{\bar{g}}$. But $w(x_{\bar{g}}, \bar{g}x_{\bar{g}}) = V(\bar{g})$. Combining this with (16) gives

$$\sum_{g=0}^{N-1} P^G(g) w(x^*, gx^*) < V(\bar{g}).$$

2. It now suffices to show that $V(g)$ is convex, since then we will have $V(\bar{g}) \leq \sum_{g=0}^{N-1} P^G(g) V(g)$. Since x_g has $w_1(x_g, gx_g) = 0$,

$$V'(g) = w_2(x_g, gx_g)(g\frac{dx_g}{dg} + x_g)$$

and

$$V''(g) = w_{12}\frac{dx_g}{dg} + w_{22} \cdot (g\frac{dx_g}{dg} + x_g) + w_2 \cdot (2\frac{dx_g}{dg} + g\frac{d^2x_g}{dg^2}) \tag{17}$$

where the arguments of w's derivatives have been dropped for clarity. We can work out dx_g/dg as follows: x_g solves $z(x_g) - gx_g/N = 0$ and therefore by the Implicit Function Theorem

$$\frac{dx_g}{dg} = \frac{x_g/N}{z'(x_g) - g/N} > 0 \tag{18}$$

since $z'(x_g) > g/N$. If $z(\cdot)$ is concave then this will increase in g, so that x_g is convex, i.e. $\frac{d^2x_g}{dg^2} > 0$. Thus, the first and third terms of (17) are positive, so that if w_{22} is close to 0, we will have V convex. (Clearly, this condition can be loosened somewhat, both in terms of the requirements on $z(\cdot)$, and in the requirement that $\frac{d^2x_g}{dg^2} > 0$.) □

Proof of Lemma 2

Proof. If $w^E(0,X)$ increases slowly in X, then $w^E(0,gx_{g+1}/N) - w^E(0,gx_g/N)$ is approximately 0 so that (9) and (8) are compatible and the lowest σ approaches 0.

If $P^E(g) \approx P^G(g)$ then the right hand side of (9) is approximately

$$\sum_{g=0}^{N-1} P^G(g) \left\{ w^E(0,gx_g/N) - w^E(0,gx_{g-1}/N) \right\}. \tag{19}$$

Now comparing the right hand side of (8), we have

$$\begin{aligned} \sum_{g=0}^{N-1} P^G(g) \left\{ w(x_g,gx_g/N) - w(\hat{x_g},gx_{g-1}/N) \right\} &\geq \sum_{g=0}^{N-1} P^G(g) \left\{ w(\hat{x_g},gx_g/N) - w(\hat{x_g},gx_{g-1}/N) \right\} \\ &\geq \sum_{g=0}^{N-1} P^G(g) \left\{ w(0,gx_g/N) - w(0,gx_{g-1}/N) \right\}, \end{aligned}$$

the first inequality by optimality of x_g, and the second by Assumption 3. But the last expression is at least as great as (19) by Assumption 6. Thus (9) and (8) can hold simultaneously.

Lastly, as N grows large, by assumption P^G approaches P^E. Then the RHS of (9), which defines the lowest possible signaling cost σ, approaches

$$\int_0^1 w^E(0,\gamma\dot{x}_\gamma) - w^E(0,\gamma\dot{x}_{\gamma-1/N})\, dF(\gamma)$$

where we have written $\dot{x}_\gamma = x_{\gamma N}$ for equilibrium good type contributions when a proportion γ of agents are good. Since x_g is increasing in g, it is continuous almost everywhere, and so is $\dot{x}_\gamma$. Similarly, since w^E is increasing in its second argument, it is continuous almost everywhere in this argument. Thus $w^E(0,\gamma\dot{x}_\gamma) \to w^E(0,\gamma\dot{x}_{\gamma-1/N})$ almost everywhere on $\gamma \in [0,1]$ and as F is continuous, the expression above goes to 0. □

Proof of Proposition 3

Proof. We take each case in turn.

1. Exclusion may still occur even with an anonymous signaling institution, since, after the number of good types is revealed, it might maximize social welfare to exclude a fraction of the players. However, if w_3 is close enough to 0, then we have for all $g \in \{0,...,N-1\}$, and $k \in \{0,...,N\}$:

$$w(x_g, gx_g/N, 1) \geq \frac{k}{N} \sum_{j=0}^{g} B_k(j) w(y_{k,g}, jy_{k,g}/N, k/N)$$

where k is the number of included players, k/N is the ex ante probability of being included, $B_k(j)$ is the probability of having j other good types from a total of $k-1$ other players who each have probability $g/(N-1)$ of being a good type, and $y_{k,g}$ is the largest equilibrium contribution under these circumstances. Proof: by assumption, $w(x,X,P) \approx w(x,X,1)$ for all x,X.Then $y_{k,g}$ is a maximizer of w given the distribution of other included good types. Clearly this is bounded above by g; then, as discussed in subsection 4.1, by increasing differences we have $y_{k,g} \leq x_g$. But then for all $j \in \{0,...,g\}$, $w(y_{k,g}, jy_{k,g}/N, k/N) \approx w(y_{k,g}, jy_{k,g}/N, 1) \leq w(y_{k,g}, gx_g/N, 1) \leq w(x_g, gx_g/N, 1)$, the first inequality by Assumption 1, the second by optimality of x_g.

In this case, then, anonymous signaling gives welfare of

$$\sum_{g=0}^{N-1} P^G(g) w(x_g, gx_g/N, 1) - \sigma, \tag{20}$$

with x_g defined as before. As before the cheapest signaling cost σ satisfies(9).

On the other hand, public signaling gives welfare of

$$\sum_{g=0}^{N-1} P^G(g) w(\tilde{x}_g, g\tilde{x}_g/N, (g+1)/N) - \sigma_X. \tag{21}$$

Here σ_X is the signaling cost when exclusion is possible. We define $\tilde{x}_g$ more formally as the

largest $x \in [0,1]$ solving the system

$$x \in \arg\max \quad w(x, X, (g+1)/N)$$
$$X = gx/N$$

Now when w_3 is small, $w(x, X, (g+1)/N) \approx w(x, X, 1)$; then by continuity of w and the Maximum Theorem, $\tilde{x}_g \approx x_g$. The first terms of (20) and (21) thus are approximately the same value: the benefit from exclusion is negligible. The proof then follows from comparing the two signaling costs:

$$\begin{aligned} \sigma_X &\geq \sum_{g=0}^{N-1} P^E(g) w^E(0, g\tilde{x}_g/N, (g+1)/N) \\ \sigma &\geq \sum_{g=0}^{N-1} P^E(g) \left\{ w^E(0, gx_{g+1}/N, 1) - w^E(0, gx_g/N, 1) \right\} \end{aligned}$$

from (10)and (9); and since $w^E(0, g\tilde{x}_g/N, (g+1)/N) \geq w^E(0, g\tilde{x}_g/N, 1)$ and $x_g \approx \tilde{x}_g$ it is easy to see that the lowest σ is less than the lowest possible σ_X.

2. Suppose $P^E \approx P^G$ and

$$w^E(0, g\tilde{x}_g/N, (g+1)/N) \approx w(\tilde{x}_g, g\tilde{x}_g/N, (g+1)/N) \tag{22}$$

for all g. [32] Then, substituting into (10), the lowest signaling cost in a public institution is close to

$$\sum_{g=0}^{N-1} P^G(g) w(\tilde{x}_g, g\tilde{x}_g/N, (g+1)/N)$$

and this is exactly the expression for good type welfare in the main game. Thus welfare in a public signaling institution is close to 0. On the other hand, the lowest signaling cost in an

[32] The latter condition requires that $w(x, X, P)$ does not vary very much in x, since $w^E(0, g\tilde{x}_g/N, (g+1)/N) < w(0, g\tilde{x}_g/N, (g+1)/N)$ by Assumption 6. Technically, we are taking two series of functions $\{w^{Ei}\}_{i=1}^{\infty}$ and $\{w^i\}_{i=1}^{\infty}$, so that $w^{Ei}(0, g\tilde{x}_g/N, (g+1)/N) \nearrow w^i(0, g\tilde{x}_g/N, (g+1)/N)$ for all g, and $w^i(0, g\tilde{x}_g/N, (g+1)/N) \nearrow w^i(\tilde{x}_g, g\tilde{x}_g/N, (g+1)/N)$ for all g. This can be done while holding (23) constant in i.

anonymous institution is again

$$\sum_{g=0}^{N-1} P^G(g) w(x_g, gx_g, 1) - \sum_{g=0}^{N-1} P^G(g) \left\{ w^E(0, gx_g/N, 1) - w^E(0, gx_{g-1}/N, 1) \right\}$$

where the second term comes from substituting P^G for P^E in(9), and this is strictly positive. Now (22) holds for $g = N-1$, and also $x_g = \tilde{x}_g$for $g = N-1$. Furthermore at . Thus, the first term and the first part of the second term add up to approximately zero, so that good type welfare is

$$\sum_{g=0}^{N-1} P^G(g) w^E(0, gx_{g-1}/N, 1) \tag{23}$$

which is strictly positive. □

Proof of Lemma 4 part (6)

Proof. After the revealed minigame, since there is pooling, average contributions are $\bar{g}x^*$ while after the anonymous minigame, average contributions are $E_g g x_g$, with the expectation being taken over the number of good types. First we show $x^* \leq x_{\bar{g}}$. By concavity of w_1 in its second argument, we have $w_1(x^*, \bar{g}x^*) \geq \sum_{g=0}^{N-1} P^G(g) w_1(x^*, gx^*) = 0$. But then we must have $x^* \leq x_{\bar{g}}$, since $(x^*, \bar{g}x^*)$ is within the upper level set where $w_1 \geq 0$, and $x_{\bar{g}}$ is the largest value such that $w_1(x_{\bar{g}}, \bar{g}x_{\bar{g}}) = 0$, or if there is no such larger value, $x_{\bar{g}} = 1$. Next, examining (18) shows that when z is concave, x_g is strictly convex in g. If so, then gx_g is also convex in g so that $E_g g x_g > \bar{g}x_{\bar{g}}$. (Indeed, this condition can be weakened to a requirement that $z''(x_g)\frac{dx_g}{dg} < 1/N$.) □

Tables

References

Acemoglu, D.: 2007, Incentives in markets, firms and governments, *Journal of Law, Economics and Organization* pp. 1–34.
URL: *http://0-jleo.oxfordjournals.org.serlib0.essex.ac.uk/cgi/reprint/ewm055v1.pdf*

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

Andreoni, J. and Petrie, R.: 2004b, Public goods experiments without confidentiality: a glimpse into fund-raising, *Journal of Public Economics* **88**(7-8), 1605–1623.

Bird, R. B. and Smith, E. A.: 2005, Signaling theory, strategic interaction, and symbolic capital, *Current Anthropology* **46**(2), 221–248.
URL: *http://www.journals.uchicago.edu/doi/abs/10.1086/427115*

Carman, K.: 2003, Social influences and the private provision of public goods: Evidence from charitable contributions in the workplace, *Manuscript, Stanford University* .

Cooper, R., DeJong, D., Forsythe, R. and Ross, T.: 1994, 7 ALTERNATIVE INSTITUTIONS FOR RESOLVING COORDINATION PROBLEMS: EXPERIMENTAL EVIDENCE ON FORWARD INDUCTION AND PREPLAY, *Problems of coordination in economic activity* p. 129.

Ellison, G.: 1994, Cooperation in the prisoner's dilemma with anonymous random matching, *Review of Economic Studies* **61**(3), 567–88.

Farrell, J. and Rabin, M.: 1996, Pre-game communication, *The Journal of Economic Perspectives* **10**(3), 103–118.

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

Fehr and Schmidt: ..., ..., ...

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

Fudenberg, D. and Tirole, J.: 1991, *Game Theory*, Mit Press.

Hagen, E. H. and Bryant, G. A.: 2003, MUSIC AND DANCE AS a COALITION SIGNALING SYSTEM, *Human Nature* **14**(1), 21–51.

Harrison, G.: 2002, Introduction to experimental economics, *At http://dmsweb. badm. sc. edu/glenn/manila/presentations* .

Holmstrom, B.: 1999, Managerial incentive problems: A dynamic perspective, *The Review of Economic Studies* **66**(1), 169–182.
URL: *http://www.jstor.org/stable/2566954*

Holt, C. and Laury, S.: n.d., Forthcoming. Theoretical explanations of treatment effects in voluntary contributions experiments. C. Plott, V. Smith, eds, *Handbook of Experimental Economic Results* .

Joseph, N. and Alex, N.: 1972, The uniform: A sociological perspective, *American Journal of Sociology* **77**(4), 719.

Kandori, M.: 1992, Social norms and community enforcement, *Review of Economic Studies* **59**(1), 63–80.

Karlan, D. and List, J.: 2007, Does Price Matter in Charitable Giving? Evidence from a Large-Scale Natural Field Experiment, *American Economic Review* **97**(5), 1774–1793.

Kreps, D. M., Milgrom, P., Roberts, J. and Wilson, R.: 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.

Levy, G.: 2007a, Decision making in committees: Transparency, reputation, and voting rules, *The American Economic Review* **97**(1), 150–168.

Levy, G.: 2007b, Decision-Making procedures for committees of careerist experts, *American Economic Review* **97**(2), 306–310.

Levy, G. and Razin, R.: 2006, A Theory of Religion: Linking Individual Beliefs, Rituals, and Social Cohesion, *Department of Economics WP, LSE* . we are not aware of empirical papers which test for a relation between natural disasters (or economic booms) and the size of religious organizations. Th.

List, J. and Lucking-Reiley, D.: 2002, The effects of seed money and refunds on charitable giving: Experimental evidence from a university capital campaign, *Journal of Political Economy* **110**(1), 215–233.

List, J. and Rondeau, D.: 2003, The impact of challenge gifts on charitable giving: an experimental investigation, *Economics Letters* **79**(2), 153–159.

Londregan, J. and Vindigni, A.: 2006, Voting as a credible threat.
URL: *http://www.princeton.edu/ pegrad/papers/londvind.pdf*

McNeill, W. H.: 1997, *Keeping Together in Time*.

Milgrom, P. and Roberts, J.: 1990, Rationalizability, learning, and equilibrium in games with strategic complementarities, *Econometrica* **58**(6), 1255–1277. ArticleType: primary_article / Full publication date: Nov., 1990 / Copyright © 1990 The Econometric

Society.

URL: *http://www.jstor.org/stable/2938316*

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

Ostrom, E., Walker, J. and Gardner, R.: 1992, Covenants with and without a sword: Self-Governance is possible, *The American Political Science Review* **86**(2), 404–417.

Prat, A.: 2005, The wrong kind of transparency, *The American Economic Review* **95**(3), 862–877.

Rafaeli, A. and Pratt, M. G.: 1993, Tailored meanings: On the meaning and impact of organizational dress, *The Academy of Management Review* **18**(1), 32–55.

Reinstein, D. and Riener, G.: 2009, Reputation and influence in charitable giving: An experiment.

Ruffle, B. J. and Sosis, R. H.: 2003, *Does it Pay to Pray? Evaluating the Economic Return to Religious Ritual*, SSRN. SSRN eLibrary.

URL: *http://ssrn.com/paper=441285*

Schram, A.: 2000, Sorting out the seeking: The economics of individual motivations, *Public Choice* **103**(3), 231–258.

Simpson, B. and Willer, R.: 2008, Altruism and Indirect Reciprocity: The Interaction of Person and Situation in Prosocial Behavior, *Social Psychology Quarterly* **71**(1), 37–52.

Smirnov, O. and Fowler, J. H.: 2007, Policy-Motivated parties in dynamic political competition, *Journal of Theoretical Politics* **19**(1), 9.

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.

Solzhenitsyn, A. I.: 1997, *The Gulag Archipelago, 1918-1956: An Experiment in Literary Investigation*, Basic Books.

Sosis, R. and Ruffle, B. J.: 2003, Religious ritual and cooperation: Testing for a relationship on israeli religious and secular kibbutzim, *Current Anthropology* **44**(5), 713–722.

Stigler, G. J.: 1972, Economic competition and political competition, *Public Choice* **13**(1), 91–106.
URL: *http://dx.doi.org/10.1007/BF01718854*

Vegetius, M. D.: 2004, *Epitoma rei militaris*, Clarendon Press New York; Tokyo: Oxford University Press, Oxford.

Watson, R. I.: 1973, Investigation into deindividuation using a cross-cultural survey technique, *Journal of Personality and Social Psychology* **25**(3), 342–5. PMID: 4705668.
URL: *http://www.ncbi.nlm.nih.gov/pubmed/4705668*

Weber, M.: 1946, *The Protestant Sects and the Spirit of Capitalism*, Oxford, New York.