Received: 20 May 2020 | Revised: 20 June 2021 | Accepted: 31 July 2021

DOI: 10.1111/jpet.12540

ORIGINAL ARTICLE

JOURNAL OF PUBLIC ECONOMIC THEORY WILEY

The whip and the Bible: Punishment versus internalization

Rohan Dutta[1] | **David K. Levine**[2] | **Salvatore Modica**[3]

[1]Department of Economics, McGill University, Montreal, Canada

[2]Department of Economics, EUI and WUSTL, Florence/St. Louis, Firenze, Italy

[3]Dipartimento SEAS, Universitá di Palermo, Palermo, Italy

Correspondence
David K. Levine, Department of Economics, EUI and WUSTL, Via della Piazzuola 43, Florence/St. Louis, Firenze I-50133, Italy.
Email: david@dklevine.com

Funding information
European University Institute, Grant/Award Number: research council; Ministero dell'Istruzione, dell'Università e della Ricerca, Grant/Award Number: PRIN 20103S5RN3

Abstract

A variety of experimental and empirical research indicate that prosocial behavior is important for economic success. There are two sources of prosocial behavior: incentives and preferences. The latter, the willingness of individuals to "do their bit" for the group, we refer to as internalization, because we view it as something that a group can influence by appropriate investment. This implies that there is a trade-off between using incentives and internalization to encourage prosocial behavior. By examining this trade-off we shed light on the connection between social norms observed inside the laboratory and those observed outside in the field. For example, we show that a higher value of cooperation outside the laboratory may lower the use of incentives inside the laboratory even as it increases their usage outside. As an application we show that the model calibrated to experimental data makes reasonable out-of-sample quantitative forecasts.

"It is not from the benevolence of the butcher, the brewer, or the baker that we expect our dinner" (Adam Smith).
"Teach self-denial and make its practice pleasure, and you can create for the world a destiny more sublime that ever issued from the brain of the wildest dreamer" (Sir Walter Scott).

1 | INTRODUCTION

A variety of experimental and empirical research indicate that prosocial behavior is important for economic success. There are two sources of prosocial behavior: incentives and preferences. People may behave prosocially because failure to do so may result in punishment by others (the whip). But even in the absence of incentives people may behave prosocially out of ethical considerations that determine their preferences (the bible): we refer to this as internalization. There is evidence for both types of behavior. For example, people make altruistic choices in double-blind treatments in the laboratory where there is no possibility of punishment or reward. The pages of history are filled with tales of great individual sacrifices for the common good. On the other hand people are not always guided by societal needs, which is why rewards and punishments exist: we do have both murderers and prisons. Here we develop a theory of group behavior in which both sources of prosocial behavior coexist and are endogenously determined. The questions we address are when we are likely to see incentives rather than internalization, whether they are complements or substitutes, and what the implications are for economic problems and empirical research.

Unlike existing models in which the prosocial behavior of some people is exogenous, we allow groups to make costly investments in producing it. We do not think there is any mystery in this. Prosocial behavior is learned and taught: by our parents, in school, and by our peers. Internalization in this view is an investment by society in changing preferences. Like social norms themselves we view this investment as endogenous and functional and ask how a group making collective decisions optimally invests in internalization. The members who as a result behave prosocially we call *acolytes*; the others will be referred to as *opportunists*. Higher investment in internalization increases the fraction of acolytes in society.

A key feature of our theory is that internalization is not possible on a case by case basis while punishment is. That is, either prosocial behavior is internalized or is not, and being in a particular environment does not have any impact on this. By contrast, incentives can be adapted to circumstances: there is no reason that the same incentive system should be used by a business firm as by a political party. Consequently the level of internalization is determined by the most important problems faced by a group. This poses issues for inferring behavior in the large from behavior in the small. Inside the laboratory, for example, we expect internalization to be exogenous but punishment endogenous. What does behavior observed in the laboratory then tell us about behavior outside the laboratory where both internalization and punishment are endogenous?.

Our theory combines several standard elements. We follow the ethical voter literature (see particularly Feddersen & Sandroni, 2006) and the experimental literature on warm-glow giving (see particularly Andreoni, 1990; Palfrey & Prisbrey, 1997) in assuming that each individual has a probability of being an acolyte, who loses utility for failing to do his/her social duty. Second, we follow a long empirical literature—in particular Coase and Ostrom (1990)—that argues that groups are good at providing incentives to their members to achieve group objectives, effectively solving mechanism design problems. The type of incentives we study are punishments as in Ostrom (1990) and Fehr and Gächter (2000). These might be social punishments such as ostracism, or even monetary punishments such as fines. We model monitoring following Levine and Modica (2016): this model has been used by Levine and Modica (2017) to study lobbying groups and by Levine and Mattozzi (2020) to study political parties. In this model there is a noisy signal of individual behavior and the possibility of imposing punishments based on these signals.

In the present setting we distinguish between the primary problem faced by a group, a stylized public good problem where the group determines the fraction of acolytes in society (at a cost), and the secondary problem, which in our applications is a laboratory experiment. The latter is much less likely or much less frequent—hence plays little role in determining the investment in acolytes. The key point is that in the secondary the fraction of acolytes is predetermined.

There are several takeaways from the theory. The first is that internalization can have large effects by complementing punishment. This is especially the case when it is difficult to provide incentives to monitors: because acolytes are willing to accept small costs to engage in honest monitoring this can be leveraged to provide incentives through punishment. The second is an important difference between the primary and the secondary problem. In the secondary problem there can be "excess" internalization—that is, it may be possible to achieve the first best without any monitoring cost simply by having acolytes engage in production. This cannot happen in the primary problem.

One of the key issues we examine is how changing the primary affects the solution of the secondary. Consider increasing the value of the public good in the primary. This will generally increase internalization, and this will spill over into the secondary. Hence if we observe societies with different primaries and compare the same secondary, for example, in a laboratory experiment, we will observe different outcomes. In particular, the level of punishment in the secondary is not monotone in the value of the public good in the primary: as the value of the public good in the primary goes up and so does internalization, in the secondary we will observe little punishment for low values as there are few acolytes willing to punish; then punishment will go up with value as more acolytes are available; and for still higher value in the primary, punishment should decrease in the secondary as the burden of production is born by acolytes. In contrast in the primary increasing the value of cooperation can never lower the level of punishment - because the fraction of acolytes in the primary is chosen optimally so there cannot be "too many" of them. Hence in experiments, since the fraction of acolytes is exogenous, the observed level of punishment need not be related to norms and incentives to cooperate in the society at large. In other words, we must be careful in inferring direct conclusions on the strength of norms in society solely based on analysis of experiments.

As specific applications to the secondary problem we engage in a quantitative calibration in laboratory experiments similar to that in the behavioral economics literature - see in particular Levine (1986) and Fehr and Schmidt (1999). Our model says that inside the lab there is a number of acolytes determined by a primary problem solved outside the lab, and that the group attempts to solve a mechanism design problem inside the lab. We first consider the classical public good experiment with punishment analyzed by Fehr and Gächter (2000), as this is similar to the type of public goods game in our basic model. In that experiment it is observed that while the average contribution is very low when there is no punishment, roughly half of the group are willing to bear the cost of punishment and by doing so induce substantial contributions. We show that this result is well explained by a simple calibration of our model.

Second, we examine dictator and ultimatum experiments. These are not ideal from our point of view since it is not entirely clear what the underlying mechanism design problem is, but the experiments are the only ones where substantial cross-cultural data is available. We observe that risk aversion will create a mechanism design problem in which there is demand for "fairness" and that several other considerations point to a social objective function of this type. Using that idea we obtain results of dictator giving quantitatively consistent with the Fehr and Gächter (2000) public goods data. We also give a reasonable out-of-sample quantitative

explanation of the ultimatum bargaining data of Duffy and Feltovich (1999). In both Fehr and Gächter (2000) and the ultimatum data there is evidence both that the participants are trying to solve a mechanism design problem and that it takes time to do it. Without communication and trying by trial and error to establish a social norm, we do not find this surprising.

Our final application is the cross-cultural ultimatum data from Henrich et al. (2001). Here we have substantial cross country variation in the value of the public good in the primary. Our theory predicts that when this is low we see very bad offers and few rejections. In the middle range we should see good offers and substantial rejections and this will be insensitive to variation in the value of the public good. Finally at the upper end offers will be very good and rejections very few again. This is indeed what we find in the data, and indeed we are able to give a reasonable out-of-sample quantitative explanation of the Henrich et al. (2001) ultimatum data.

We emphasize that our goal in this paper is a kind of "proof of concept:" can a simple model of mechanism design with acolytes capture aggregate behavior in some relevant experimental data?.[1]

2 | ECONOMIC ENVIRONMENT

We study an organized group with many members engaging in a representative producer-recipient-monitor interaction. There are two possible states: the *primary* state $s = 1$ which we interpret as the "normal" state of affairs and the *secondary* state $s = 2$ which is much less likely, for instance a laboratory experiment. After the state is known the producer chooses an amount $x_s \geq 0$ to produce at unit marginal cost. Output represents a public good providing a social benefit to the recipient of $V_s f(x_s)$ where $V_s > 0$ is a measure of the value of the public good and f is smooth and strictly differentiably increasing and concave[2] with $V_s f'(\infty) < 1$. The *first best*, that is the x_s which maximizes $V_s f(x_s) - x_s$, is then the unique solution of $V_s f'(x_s) = 1$, or 0 if $V_s f'(0) \leq 1$.

The effect of any individual member in a large group on average output is negligible, so there is a severe free-rider problem. We have modeled this by separating the recipient from the producer. Hence a selfish producer would prefer not to produce at all. We are going to assume that peer pressure can be used to provide producer incentives: production can be monitored and those who fail to produce can be punished. Specifically, in state s the group may establish an output quota y_s and generate a noisy signal $z_s \in \{0, 1\}$ about whether the producer respected the quota (i.e., produced at least y_s), where 0 means "good, respected the quota" and 1 means "bad, failed to respect the quota." If the quota is not satisfied the signal takes on the value 1 with probability one, and if it is satisfied it takes on the value 1 with probability $\pi_s < 1$. This simple stark signal technology works well in our quantitative analysis and our qualitative analysis is robust to more general error processes.

The production signal is observed by an anonymous monitor who, if the signal about the producer is bad, chooses whether or not to transmit it. If a bad signal is transmitted the group imposes an endogenous utility penalty $P_s \geq 0$ on the producer. This may be in the form of ostracism or some other social penalty. This punishment is costly for the monitor, who bears a

[1]We emphasize that we do not try to explain individual behavior: equilibrium models as a rule do not do a good job with individual behavior and alternatives such as quantal response models are generally used to analyze the behavior of individuals. See, for example, Levine and Palfrey (2007).
[2]That is $f' > 0$ and $f'' < 0$.

proportionate cost of $\psi_s P_s$ where $\psi_s > 0$. Notice that since $\psi_s > 0$ selfish monitors will never transmit the signal.

As we indicated in the introduction there are two types of group members: *acolytes* and *opportunists*. Types determine preferences in the sense we will specify shortly. They are private and drawn independently. The probability $0 \leq \phi \leq 1$ of being an acolyte is endogenous and applies to both states. It is chosen in the primary state, that is, it is targeted towards the "normal" state of affairs.

A *social norm in states* consists of an output quota y_s, an output target for acolytes $Y_s \geq y_s$, and a punishment level P_s. The group faces a mechanism design problem. In the primary problem this consists a choice of ϕ and a choice of social norm. In the secondary problem ϕ is imported from the solution of the primary problem and a social norm is chosen accordingly.

A social norm is only meaningful if group members are willing to adhere to it. In this context that means that it is incentive compatible for acolytes to produce Y_s, for opportunists to produce y_s and for acolytes to transmit a bad signal. Incentive compatibility is defined with respect to an internalization penalty $\gamma > 0$: any acolyte who does not follow the social norm suffers a penalty of that amount. In other words, an acolyte producer who fails to hit the output target Y_s or an acolyte monitor who fails to transmit a bad signal loses utility γ. This can be interpreted as guilt for violating the social norm. Opportunists suffer no penalty. When a social norm is followed in state s each type of producer meets the output quota, therefore expected output is $x_s = (1 - \phi)y_s + \phi Y_s$. The probability of generating a bad signal is therefore π_s and the probability that this signal is transmitted and the producer is punished is equal to the probability that the monitor is an acolyte, that is ϕ; the social cost of this punishment is the cost to the producer plus the cost to the monitor $P_s + \psi_s P_s$. Therefore the social utility under the incentive compatible norm (y_s, Y_s, P_s) given ϕ is

$$\begin{aligned} U_s &= V_s f((1-\phi)y_s + \phi Y_s) - ((1-\phi)y_s + \phi Y_s) - \phi\pi_s(1+\psi_s)P_s \\ &= V_s f(x_s) - x_s - \phi\pi_s(1+\psi_s)P_s. \end{aligned}$$

The last term is the extra cost generated by the need to solve the free-rider problem. Before the realization of the state the group invests in indoctrination: the greater the investment in indoctrination the greater the probability ϕ that a group member will be an acolyte.[3] Such investment is costly: the social cost is $H\phi$. The choice of ϕ is made as indicated in the primary state hence we attribute the investment cost to that state. In particular the objective function in the primary state is $U_1 - H\phi$, and $\phi, (y_1, Y_1, P_1)$ is chosen to maximize this. In the secondary state ϕ is taken as given and (y_2, Y_2, P_2) is chosen to maximize U_2.

2.1 | The incentive constraints

In the sequel the state will be often clear from context: in these cases we will omit the state subscript. To more clearly understand the mechanism design problem and the model it is useful to derive the incentive constraints. In the production problem the probability of being punished is equal to the probability that the monitor is an acolyte times the probability of a bad

[3]In other words the penalty suffered by acolytes γ is exogenous but the fraction of acolytes ϕ is endogenous. This is discussed below in Section 3.6.

signal. Hence for an opportunist the cost of meeting the target y is $y + \phi\pi P$; the best alternative is to produce zero, at cost ϕP. Therefore the incentive constraint for an opportunist is $y + \phi\pi P \leq \phi P$ or $y \leq \phi(1 - \pi)P$.

Note that whenever it is incentive compatible for an opportunist to produce y it is incentive compatible for an acolyte to produce up to $y + \gamma$. We define φ by $\varphi = (Y - y)/\gamma$ so that $Y = y + \varphi\gamma$. Then the above says that a norm (y, Y, P) is incentive compatible for both types of producers if and only if $y \leq \phi(1 - \pi)P$ and $0 \leq \varphi \leq 1$. Then a norm (y, Y, P) can be equivalently expressed as (y, φ, P); we shall most often use the latter form. Using this notation we can write $x = (1 - \phi)y + \phi Y = y + \phi\varphi\gamma$, and

$$U = Vf(y + \phi\varphi\gamma) - (y + \phi\varphi\gamma) - \phi\pi(1 + \psi)P.$$

Monitoring as we have indicated can only be carried out by acolytes. For them incentive compatibility requires that the private cost of monitoring not exceed the internalization penalty. This results in the monitoring incentive compatibility constraint $\psi P \leq \gamma$. It will be easily seen that the producers constraint $y \leq \phi(1 - \pi)P$ binds; thus the monitoring constraint will be $y \leq \phi(1 - \pi)\gamma/\psi$.

Finally, it will be convenient to have a notation for the upper bound on the output the group can produce. It follows from $y \leq \phi(1 - \pi)\gamma/\psi$ that it must be $x \leq \phi(1 - \pi)\gamma/\psi + \phi\varphi\gamma$. Since $\varphi, \phi \leq 1$ we get

$$x \leq (1 - \pi)\gamma/\psi + \gamma \equiv \chi.$$

2.2 | Preliminaries

We will repeatedly solve optimization problems equivalent to $\max_x f(x) - \mu x$ subject to $\underline{x} \leq x \leq \underline{x} + X$. Since $f'(x)$ is by assumption strictly decreasing this has a unique solution x^* given by the solution to the first order condition $f'(x) = \mu$ if this is feasible, and lying on the relevant boundary if it is not. We can conveniently write the solution as $x^* = g(\mu, \underline{x}, X)$ where g is $[f']^{-1}(\mu)$ truncated to satisfy the constraints, that is

$$g(\mu, \underline{x}, X) = \begin{cases} \underline{x} + X & \text{if } [f']^{-1}(\mu) > \underline{x} + X \\ [f']^{-1}(\mu) & \text{if } \underline{x} + X \geq [f']^{-1}(\mu) \geq \underline{x} \\ \underline{x} & \text{if } \underline{x} > [f']^{-1}(\mu). \end{cases}$$

In our applications we will be solving problems of the form $\max_\theta Vf(a + b\theta) - c\theta$ subject to $0 \leq \theta \leq \Theta$ which by a transformation is equivalent to the simpler form above so that the solution θ^* can also be expressed in terms of the function g. This expression together with a summary of the properties of the function g is as follows.

Lemma 1. *The function $g(\mu, \underline{x}, X)$ is continuous and increasing in $\underline{x}, X$.*[4] *It satisfies $\underline{x} \leq g(\mu, \underline{x}, X) \leq \underline{x} + X$ and for $\underline{x} < g(\mu, \underline{x}, X) < \underline{x} + X$ it is smooth and strictly*

[4]For brevity increasing and decreasing without qualification always mean weakly so.

decreasing in μ. *The solution to* $\max Vf(a + b\theta) - c\theta$ *subject to* $0 \leq \theta \leq \Theta$ *is unique and given by* $\theta^* = (1/b)(g((1/V)(c/b), a, b\Theta) - a)$.

The proof of all results can be found in Appendix A.

3 | OPTIMAL SOCIAL NORMS

We first analyze the optimal social norm (y^*, φ^*, P^*) for a given value of ϕ. This gives the solution of the secondary problem where ϕ is in fact fixed, and will enable us to solve the primary problem for the optimal value of ϕ. For ease of reading we will continue to omit the state subscript.

A key idea in the choice of an optimal social norm is encapsulated in the *marginal cost of monitoring*

$$M \equiv (1 + \psi)\frac{\pi}{1 - \pi}. \tag{1}$$

As we will see this measures the marginal cost of increasing output y produced by opportunists, arising from the need to punish them. It consists of two parts: the first $1 + \psi$ is the social cost of punishment, the second $\pi/(1 - \pi)$ measures the difficulty of monitoring. Notice that the numerator π plays a key role: it measure the amount of punishment that takes place on the equilibrium path - that is erroneous punishment.

Theorem 1 (Optimum in the secondary). *At the optimal solution if* $\phi = 0$ *then* $y^* = 0$ *and* φ, P *do not matter. When* $\phi > 0$ *then* $\varphi^* = (1/(\phi\gamma))g(1/V, 0, \phi\gamma)$, *and*

1. *If* $Vf'(\phi\gamma) \leq 1$ *the optimal solution is first best with* $y^* = P^* = 0$.
2. *If* $Vf'(\phi\gamma) > 1$ *the solution is* second best *with* $\varphi^* = 1$,

$$y^* = g\left(\frac{1 + M}{V}, \gamma\phi, \frac{(1 - \pi)\phi\gamma}{\psi}\right) - \gamma\phi,$$

and the producers' constraint binding:

$$P^* = \frac{y^*}{\phi(1 - \pi)}.$$

In this case there are two subregimes depending on whether y^* *is at the corner where opportunists are not used to produce output: from the definition of* g *this occurs where* $Vf'(\phi\gamma) = 1 + M$.

In addition maximized utility is concave and increasing in ϕ. *Finally,* $y^* \leq (1 - \pi)\phi\gamma/\psi$.

The theorem says that there are three regimes. If $Vf'(\phi\gamma) \leq 1$ we should use just the acolytes to provide output; the reason is that there is no monitoring cost associated with their providing up to $\phi\gamma$ of output, and it is enough in the sense that the first best of maximizing $Vf(\phi\varphi\gamma) - \phi\varphi\gamma$ is achieved for $\varphi^* < 1$. If $Vf'(\phi\gamma) > 1$ all acolytes should be used to produce output, so $\varphi^* = 1$. There are two subregimes. If $1 < Vf'(\phi\gamma) \leq 1 + M$ it is not worth using opportunists to produce output. If $Vf'(\phi\gamma) > 1 + M$ it is optimal to provide incentives to non-acolytes to produce too. In other words, punishment serves as a costly backstop technology to making use of acolytes who have internalized the social norm.

It is worth stressing the fact that the extent of optimal punishment, given the fraction of acolytes, is proportional to the optimal quota.

3.1 | Comparative statics of the secondary

Since ϕ is fixed in the secondary we are now in a position to describe the comparative statics. We focus on the optimal norm in the case $Vf'(\phi\gamma) > 1$: if $Vf'(\phi\gamma) \leq 1$ we can attain the first best in the secondary simply by having acolytes produce.

Corollary 1. *If $Vf'(\phi\gamma) > 1$ then $\varphi^* = 1$ and total output $x^* = y^* + \phi\gamma$ is increasing in V, ϕ and decreasing in π, ψ. Define $\hat{\phi}$ by $Vf'(\hat{\phi}\chi) = 1 + M$. For $\phi < \hat{\phi}$ the optimal quota y^* and punishment P^* are increasing in ϕ and for $\phi > \hat{\phi}$ they are decreasing.*

Note that the last part asserts the non-monotonicity of punishment in the value of the public good (on which we will expand in Section 3.5). To see what is going on observe that from Theorem 1 for $Vf'(\phi\gamma) > 1$ the quota y^* maximizes $Vf(y + \phi\gamma) - (1 + M)y$ on $0 \leq y \leq \phi(1 - \pi)\gamma/\psi$; so $y + \phi\gamma \leq \phi(1 - \pi)\gamma/\psi + \phi\gamma = \phi\chi$. For $\phi < \hat{\phi}$ it is $Vf'(\phi\chi) > 1 + M$, so y^* is at the upper bound $\phi(\chi - \gamma)$, increasing in ϕ; for $\phi > \hat{\phi}$ the optimal y^* is given by the first order condition $Vf'(y^* + \phi\gamma) = 1 + M$ and so higher ϕ calls for lower y^*.

3.2 | Solution of the primary problem

To solve the primary we need to optimally choose ϕ and a corresponding social norm for the primary state $s = 1$. Since by Theorem 1 we already know the optimal choice of social norm for any state s and any ϕ, we simply need to find the optimal ϕ for the primary state $s = 1$. Again, since we are dealing entirely with one state, we omit state subscripts.

Since P must be chosen optimally from Theorem 1 it must satisfy $P = y/(\phi(1 - \pi))$. Substituting this into the objective function for the primary we see that the objective may be written as

$$W \equiv U - H\phi = Vf(y + \phi\varphi\gamma) - (y + \phi\varphi\gamma) - My - H\phi.$$

Moreover, also from Theorem 1y y must be chosen optimally, so according to that theorem must satisfy the constraint $y \leq (1 - \pi)\phi\gamma/\psi$; and of course $\varphi \in [0, 1]$. Hence we can solve the primary by maximizing W with respect to ϕ, y, φ subject to these two constraints.

Lemma 2. *The optimal primary social mechanism has* $\varphi^* = 1$.

Theorem 2 (Optimum in the primary). *If* $H < \gamma M$ *then* $\phi^* = (1/\gamma)g((1/V)(1 + H/\gamma), 0, \gamma)$ *and the optimal quota is*

$$y^* = g\left(\frac{1+M}{V}, \gamma, \chi - \gamma\right) - \gamma$$

which is equal to zero if $\phi^* < 1$.

If $H > \gamma M$ *then*

$$\phi^* = \frac{1}{\chi} g\left(\frac{1}{V}\frac{\chi + (1+\psi)\pi\gamma/\psi + H}{\chi}, 0, \chi\right)$$

and $y^* = (1-\pi)\phi^*\gamma/\psi$.

The theorem has two cases. If $H < \gamma M$ then acolytes are cheap and punishment expensive, and output is produced solely by acolytes (if there are opportunists, i.e., $\phi < 1$, they will not produce). If $H > \gamma M$ then acolytes are expensive and punishment cheap so as much punishment as is possible should be used to get output from opportunists (y^* is at its upper bound).

3.3 | Comparative statics of the primary

Corollary 2. *In the primary problem internalization* ϕ^*, *the production quota* y^*, *total output* $x^* = y^* + \phi^*\gamma$ *and punishment* P^* *are increasing in* V. *Total output* x^* *is increasing in* γ *and decreasing in* π. *If* $H > \gamma M$, *punishment* P^* *is constant in* V, *and for* $0 < \phi^* < 1$ *the optimal* ϕ^*, y^*, x^* *strictly increase.*

We stress the case $H > \gamma M$ because we argue that in applications it is the more relevant one. In that case the punishment $\psi P^* = \gamma$ is at its upper bound hence independent of V.

3.4 | Lessons learned

There are several takeaways from this analysis. First, internalization is essential for monitors: in this model no monitoring can take place without internalization because monitoring is costly and monitors cannot be monitored.[5] It is a ubiquitous problem in mechanism design that getting people to tell the truth about others is problematic. If monitors have incentive to lie, for example, because punishment either is costly or beneficial to them, and they can be identified, then it is possible to make them indifferent by punishing them based on their reports. However, this provides weak incentives for truth-telling and if monitoring itself is costly, there is no incentive to bear that cost. Even a small incentive to tell an undetectable lie can lead to

[5]Or it is prohibitively expensive to do so: see Levine and Modica (2016) for a model where monitors can be monitored.

enormous losses—and a small amount of internalization by making it strictly optimal for acolytes to tell the truth can have a big impact.[6]

The second take-away is that in this simple model there is a single variable "internalization" ϕ that links problems across states. This has also been called "publicness" and "pro-social." It plays a key role in solving the second stage problem as Theorem 1 shows. One particular implication is that if we can measure ϕ as we do below using laboratory data then it tells us something about the solution of the mechanism design problem outside the laboratory.

The role of internalization also differentiates societies. That is, societies facing different primary problems will choose different levels of internalization and this means that they will choose different solutions to secondary problems: we examine this next.

3.5 | Connecting the primary and the secondary

Here we take up the issue of how changing the importance of public goods in a society, that increasing the value V_1 in the primary, impacts on economic outcomes in both the primary and the secondary. The remaining parameters are held fixed, although they need not be equal in the primary and the secondary. As we believe that it is common to observe less than complete internalization and some degree of punishment, we focus on the case of costly acolytes: $H > \gamma M_1$. We limit attention to the case

$$V_1 > \frac{\chi_1 + (1+\psi_1)\pi_1\gamma/\psi_1 + H}{\chi_1 f'(0)} \equiv \underline{V}_1$$

since otherwise by Theorem 2 there will be no acolytes and no output in either state. Similarly we limit attention to

$$V_1 < \frac{\chi_1 + (1+\psi_1)\pi_1\gamma/\psi_1 + H}{\chi_1 f'(\chi_1)} \equiv \overline{V}_1$$

since otherwise the number of acolytes $\phi^* = 1$ and further increases in V_1 will have no impact on either state.

About the secondary we assume that $V_2 > (1+M_2)/f'(0)$ so that some output is desirable. We focus as well on secondaries that are not only relatively unlikely, but also, like laboratory experiments, with substantially lower stakes than the primary. Other examples might include politeness in greeting people and throwing trash in bins rather than littering. Specifically we assume that $V_2 < (1+M_2)/f'(\chi_2)$ which says that it is not worth carrying out punishments to the extent needed to get the highest feasible level of output.

Theorem 3. *If $H > \gamma M_1$ and $\underline{V}_1 < V_1 < \overline{V}_1$ then as V_1 increases*

(i) *Acolytes ϕ^*, output x_1^*, and the quota y_1^* all strictly increase, and punishment P_1^* is constant.*

[6]This is not a paper about monitoring technology: in addition to monitoring monitors it may be that there are several monitors whose reports can be compared. For a deeper analysis of monitoring monitors see Rahman (2012). We chose this simple technology to make the point that internalization can be essential.

If in addition $V_2 f'(\phi^*\gamma) > 1 + M_2$ *there are intermediate cutoffs* $\underline{V}_1 < V_1^m < V_1^M < \overline{V}_1$ *such that*

(ii) *for* $\underline{V}_1 < V_1 < V_1^m$ *output* x_2^* *and the quota* y_2^* *strictly increase while punishment* P_2^* *is constant. For* $V_1^m < V_1 < V_1^M$ *output* x_2^* *is constant and the quota* y_2^* *and punishment* P_2^* *strictly decrease. For* $V_1^M < V_1 < \overline{V}_1$ *output* x_2^* *strictly increases, the quota* $y_2^* = 0$, *and punishment is constant at zero.*

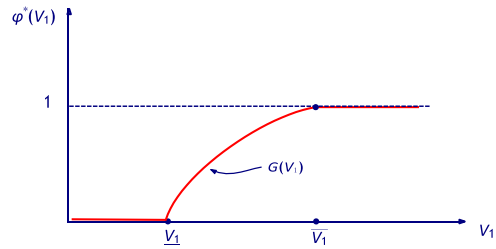
What does this result tell us about internalization? The fraction of acolytes ϕ^* is a measure of internalization, the quota y_s^* is the output induced through the use of incentives. Alternatively incentives are determined by the expected punishment conditional on a bad signal, which is $\phi^* P_s^* = y_s^*/(1 - \pi_s)$, proportional to the quota, or by the total social cost of punishment which is $(1 + \psi)\pi\phi^* P_s^* = M_s y_s^*$, also proportional to the quota.

We see then that, as V_1 increases until V_1^m is reached, internalization and incentives are complements in both problems in the sense that both strictly increase. After V_1^m is reached, in the secondary only they become substitutes. Increases in V_1 above this level continues to raise the number of acolytes from the primary, but now the marginal value of output in the secondary has dropped enough that it is optimal to take advantage of the extra production available from the additional acolytes to reduce the output quota y_2^* rather than further increasing output. By contrast in the primary we would never choose the level of internalization this high, so punishment and internalization always rise. Eventually at V_1^M the quota in the secondary has dropped to zero so no further cost reductions are possible, and output again increases. This difference in solutions between the primary and secondary problem means that we cannot reach simple and direct conclusions on the primary based on observing the secondary. In particular: if we observe little punishment in a laboratory experiment this does not imply that there is little punishment in the society at large.

Example We now illustrate the theory with a parametric example. We continue to assume $H > \gamma_1 M_1$. We take $f(x) = x - (1/(2\beta))x^2$ where we assume that $\beta > \chi_s$ so that the function is strictly increasing in the feasible region $[0, \chi_s]$. Let

$$G(V_1) = \frac{\beta}{\chi_1}\left[1 - \frac{\underline{V}_1}{V_1}\right].$$

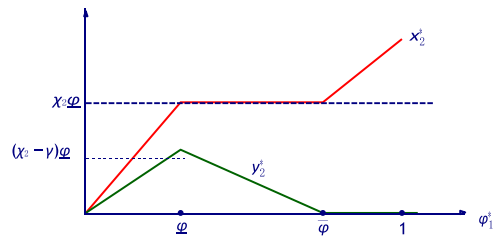
The optimal ϕ_1^* as a function of V_1 is equal to G for $\underline{V}_1 \le V_1 \le \overline{V}_1$, while $\phi_1^* = 0$ for $V_1 < \underline{V}_1$ and $\phi_1^* = 1$ for $V_1 > \overline{V}_1$. Observing that case $\underline{V}_1/\overline{V}_1 = 1 - \chi_1/\beta$, the graph of the optimal fraction of acolytes is as plotted below.



Here $y_1^* = [(1-\pi_1)\gamma/\psi_1]\phi^*$ and $x_1^* = [(1-\pi_1)\gamma/\psi_1 + \gamma]\phi^*$ are both proportional to ϕ^*. More interesting is the behavior of y_2^*, x_2^*. Assume $V_2 > 1/(1-\gamma/\beta)$ so that $\varphi_2^* = 1$. Define

$$\underline{\phi} = \frac{\beta}{\chi_2}\left[1 - \frac{1+M_2}{V_2}\right], \quad \overline{\phi} = \frac{\beta}{\gamma}\left[1 - \frac{1+M_2}{V_2}\right]$$

(recall that higher ϕ_1^* correspond to higher V_1). Then y_2^*, x_2^* as a function of ϕ_1^* are as plotted below, illustrating the force of Theorem 3:



The graph is drawn for $\overline{\phi} < 1$. However it may be that $\underline{\phi}$ is bigger than one (in which case so is $\overline{\phi}$) or that it may be that $\underline{\phi} < 1 < \overline{\phi}$, in which cases the graph must be appropriately truncated.

3.6 | Investment in acolytes

We assumed that the penalty for acolytes γ is exogenous and that investment in indoctrination change the fraction of acolytes ϕ which is endogenous. As a practical matter we think that γ is heterogeneous in the population and that investment in indoctrination shifts this distribution up. For simplicity we model this distribution as having just two types. There are two simple models of how this distribution might shift with investment: that the fraction of acolytes goes up holding fixed γ at an exogenous level or that the level of acolyte devotion γ goes up with the fraction of acolytes ϕ held fixed at an exogenous level. Here we use the former model. In earlier working paper version of this study (Dutta et al., 2017) we used the latter model. We switched to the current model for expositional reasons. The results on the secondary in Theorem 1 are the same since both γ and ϕ are held fixed there. Not surprisingly analogous with Theorem 2 in primary if ϕ is held fixed then as V_1 increases then γ^* increases. This preserves the key result Theorem 3: that as V_1 goes up in the secondary output x_2^* rises while the quota y_2^* first rises then falls. This can easily be seen from Theorem 1 where what matters for output and the quota is $\phi\gamma$ so increasing γ^* with V_1^* has the same effect as increasing ϕ^*.

4 | LABORATORY EXPERIMENTS

According to our theory laboratory experiments are a secondary problem. That is, we can be reasonably confident that internalization is determined without reference to the possibility that group members may find themselves under study by social scientists. This means ϕ is pre-determined and participants will solve the mechanism design problem posed in the laboratory

taking this as given. We now pose the question: can laboratory data be well explained this way? In other words, are data consistent with the existence of a fraction of acolytes as our theory predicts?.

We should emphasize the following. Our benchmark producer/monitor/recipient model has been chosen for illustrative purposes. There are many other mechanism design problems that might in practice constitute either the primary or secondary. For example, production might involve joint effort by several group members, there might be several monitors, output might have both public and private dimensions, and so forth. They key point is that internalization ϕ is determined in the primary and is a parameter in the secondary and that given ϕ participants solve a mechanism design problem in the laboratory. There is no need to limit our experimental analysis to the particular mechanism design problem used to illustrate the basic theory.

4.1 | Monitoring

To move to applications we need to look more closely at the monitoring technology, in particular how a wrong signal about a member's behavior may arise in the laboratory. In each case there is a social norm in the form of an output quota y, and the signal z^i on member i's behavior is about whether or not output $x^i \geq y$. There are three possible sources of noise: first, it may be that x^i is imperfectly observed, which we think is the most common interpretation. In the laboratory as a rule x^i is perfectly observed so we reject this source of noise in our applications. We turn next to two sources of noise that are relevant to the laboratory.

A second source of noise may be that the social norm y is imperfectly observed. Here π corresponds to uncertainty over the social norm. An ultimatum bargaining experiment is, for example, an unusual event, and two different members of a group may well have different interpretations of how the social norm applies. In the public goods experiments we study, in three different sessions average output ranges from 9.8 to 14.3 which might indicate that there is substantial uncertainty about what the social norm is. However this interpretation also is problematic in the sense that over time uncertainty about the social norm should diminish and over a sufficiently long period we might expect π to converge to zero. We will present evidence showing that this is unlikely to be the case.

The third source of noise can be described as "bad behavior." In assessing this it is important to recognize that what π measures is "on-path" punishment, something important in resolving the mechanism design problem and something we see both in experiments and the broader world at large. In the model acolytes are homogeneous and so are opportunists: in practice they are not, nor would anyone who has looked at raw laboratory data imagine that they are. Hence we may take an approach like in Harsanyi (1973) and consider that there is a chance of deviant preferences. Members may wake up on the wrong side of the bed, they may pursue objectives other than maximizing income, such as showing up other players, they may be bored, they may have different risk preferences and some may in fact be risk loving, or they may experiment to check that they will indeed be punished for violations of the social norm. We can think of this as a probability of an output deviation around the social norm, and deviations to lower output correspond to "bad behavior" and lead to "bad signals." Unlike confusion over social norms, "bad behavior" is a persistent source of bad signals even when players social norms are well understood and output perfectly observed. This is going to be our preferred interpretation of π in the experiments. In our applications we assume more

specifically that when the social norm involves a positive level of output or punishment the misbehavior is a symmetric deviation from the social norm so that the expected quantities calculated from the theory do not change due to the presence of misbehaving players. If players are called upon to do nothing (not produce or not punish) we assume that they do not "misbehave" with respect to taking no action.

4.2 | Learning

In assessing laboratory data we may start with the following consideration: if in fact acolytes are successfully solving a public goods problem then realized utility in the laboratory must be greater than would be obtained in the *no-punishment* mechanism where acolytes contribute but do not punish. We refer to this as *break-even*. This test is relatively easy to implement, and indeed in the literature on punishment to induce public goods contributions authors have asked exactly this question: does realized utility including the cost of punishment exceed the utility achieved from voluntary contribution to the public good without punishment?.

We should also emphasize that the laboratory is an especially difficult environment for solving mechanism design problems. Agreement over a social norm must be reached without the possibility of discussion and based on limited observation of the behavior of other participants in a small number of matches. We do not think that people instantaneously solve mechanism design problems any more than they instantaneously solve optimization problems. Hence, as is common in the study of equilibrium, we will wish to focus on later rounds after learning has taken place.[7]

Our starting point will be the classical experiment of Fehr and Gächter (2000) on the use of punishments to induce contributions to a public good. A crucial finding in that paper is that break-even is achieved only in the final rounds of 10 rounds of play. In our study below of ultimatum bargaining we find that in the standard 10 round design there is no evidence that participants have been successful in solving a mechanism design problem, although they seem to get it right when they play additional rounds. This confirms our thought that finding an optimal mechanism can occur only with substantial learning. Hence we are limited to experimental studies in which participants engage in the same game over a substantial number of periods so have ample opportunity to learn. Unfortunately this rules out many experimental studies: for example, most studies of the trust game involve one or only a few rounds, studies of ultimatum that vary the parameters faced by the participants repeat each set of parameters only a few times and so forth.

4.3 | Overview of findings

Before jumping into the details of the experimental analysis here is an overview of our findings. We study three classes of games in which the subjects are Western college students: a public goods game, the dictator game, and ultimatum bargaining. First, it appears that the probability of being an acolyte is about 50% and that it takes at least 10 rounds of play to "solve" the

[7]The literature on level-k beliefs, for example, Stahl and Wilson (1995), show clearly that equilibrium play is not a good description of the first round in the laboratory, while repeated strangers treatments often lead to equilibrium even in environments where finding equilibrium is computationally demanding, see, for example, Levine and Palfrey (2007).

mechanism design problem posed in the laboratory. The theory works well quantitatively for both the public goods problem and for dictator games. For ultimatum bargaining games the results are mixed. If the only source of the demand for fairness is risk aversion then the theory fairs poorly, but if there is substantial demand for fairness for other reasons the theory fairs well.

5 | PUBLIC GOODS AND PUNISHMENT IN THE LABORATORY

The classical experiment on the use of punishments to induce contributions to a public good is that of Fehr and Gächter (2000). They study a public goods contribution game with four players. They examine treatments both with and without the possibility of punishment. Participants choose contribution levels $0 \leq x^i \leq 20$ and receive utility $u^i = v_0 - cx^i + v\sum_{j\neq i} x^j$ where $v_0 = 20, v = 0.4$, and $c = 0.6$.

We analyze their results for the final round of 10 in the stranger treatment.[8] As indicated, we examine the final round to allow participants the chance to "learn their way" to a solution. Although Fehr and Gächter (2000) also study a partners treatment in which all 10 rounds are played with the same partners, we know from the work of Bó (2005) that we need repeated treatments of such a repeated game to observe equilibrium play. Hence we focus on the stranger treatment. We use data averaged across all three sessions.

The average contribution in the no-punishment condition is $x = 1.9$. In the punishment treatment, we will shortly describe, contributions were much higher, at $x = 12.3$. Can our theory of internalized norms possibly account for such large contributions when there is punishment? Surprisingly the answer is yes: with the costs and consequences of punishment the acolytes can be leveraged to greatly enhance contributions.

5.1 | The punishment game

We must describe how the punishment treatment works. After contributions are observed, participant i can purchase punishment points p_i^j against j. The cost of these points is equal to the number of points up to 2 points, then becomes convex.[9] As we explain later our theory does not suggest purchases greater than 2.43 so we treat the cost of punishment points as linear. Each punishment point against a participant reduce their payoff by 10%: specifically utility at the end of the punishment round is $v^i = (1 - (1/10) \cdot \min\{10, \sum_{j\neq i} p_j^i\})u^i - \sum_{j\neq i} p_i^j$, where the min avoids pushing payoff below zero.

5.2 | The mechanism design problem

As indicated, we interpret noise in the signal z^i as due to bad behavior. It then makes sense to assume that all four participants observe the same signal z^i. We also assume that if there is a

[8]In the strangers treatment the group composition is randomly changed from period to period.
[9]The cost of 3 points is 4, for example.

bad signal for any match participant all the acolytes choose a common number of punishment points which we denote by p.

To analyze the induced incentives we must recognize that several participants may have bad signals and that therefore punishers may have to split their punishments. Conditional on receiving a bad signal, let Q be the expected number of *potential punishers*, that is those who would punish provided they were acolytes. Since each individual has probability ϕ of being an acolyte, conditional on a bad signal, the expected punishment is then ϕQp. In Lemma A3 in Appendix A we show that $Q = 3(1-\pi) + \pi^2$.

For an opportunist then the utility from abiding by the social norm of y with average output x is $(1 - \pi Qp\phi/10)(v_0 - cy + 3vx)$ and from contributing zero is $(1 - Qp\phi/10)(v_0 + 3vx)$, where notice that the free rider has no punishment cost because she does not punish. Hence the incentive constraint is $(1 - \pi Qp\phi/10)(v_0 - cy + 3vx) \geq (1 - Qp\phi/10)(v_0 + 3vx)$.

Next we need to determine how much extra $Y - y$ an acolyte is willing to produce. The fact that there is an expected cost of punishing in the punishment round limits what acolytes will be able to contribute in the first. Specifically, the expected cost of punishing in the punishment round is $(1 - (1-\pi)^3)p$. Hence the extra cost that can be carried in the first period is $\gamma - (1 - (1-\pi)^3)p$. This gives $Y - y = (\gamma - (1 - (1-\pi)^3)p)/c$.

The mechanism design problem can now be stated: it is to maximize over y, Y, x, p the objective

$$W = (1 - \phi Qp\pi/10)(v_0 - cx + 3vx) - (1 - (1-\pi)^3)\phi p$$

subject to feasibility $x = y + \phi(Y - y)$, incentive compatibility for the opportunists

$$(1 - \pi Qp\phi/10)(v_0 - cy + 3vx) \geq (1 - Qp\phi/10)(v_0 + 3vx)$$

and the two incentive compatibility constraints for the acolytes:

$$p \leq \gamma, \quad Y - y = (\gamma - (1 - (1-\pi)^3)p)/c.$$

Since the objective is linear and increasing in y and the opportunistic incentive compatibility constraint is linear, it follows that the opportunists constraint must hold with equality. Solving it for x we get

$$x = \frac{(1 - \pi Qp\phi/10)c\phi(Y - y) + (1-\pi)(Qp\phi/10)v_0}{(1 - \pi Qp\phi/10)c - (1-\pi)(Qp\phi/10)3v}.$$

5.3 | Calibration: Acolytes and opportunists

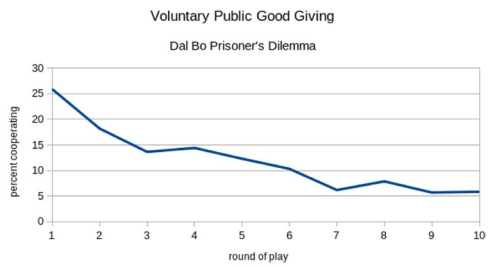
Our first goal is to determine the fraction of acolytes ϕ and the strength of their commitment γ. We do this using different data than the data we will analyze: we do this by looking at the Fehr and Gächter (2000) treatments in which there is no punishment and only voluntary contributions to the public good occur.

Quoting Fehr and Gächter (2000), the key fact is that "in the no-punishment condition of the stranger-treatment average contributions converge close to full free-riding over time."

In particular the average contribution was $x = 1.9$.[10] Moreover "we call those subjects 'free-riders' who chose ...[to contribute 0]... in more than five periods of the no-punishment... [They constitute] 53 percent in the Stranger-treatment." Adopting this definition would yield $\phi = 0.47$.

Knowing that in the stranger treatment $x = 1.9$ enables us to compute γ. The cost of average contribution $cx = 0.6 \times 1.9 = 1.14$ is also equal to the average $(1 - \phi) \times 0 + \phi\gamma$ obtained when acolytes contribute the most they are willing. From this, with $\phi = 0.47$ we get $\gamma = 2.43$.

There are issues with this analysis. First, in the no-punishment treatment the number of people contributing and the level of contributions decline over time. In particular the average over all 10 periods of the contribution is 3.7, considerably higher than in the final period, while the number of non-contributors in the final period is about 75%, considerably higher than the Fehr and Gächter (2000) estimate of 53% opportunists. The first issue is whether in fact contributions were continually declining so that if play continued longer they would continue to decline. Fehr and Gächter (2000) do not provide round-by-round data about contributions, but this type of decline is typical for voluntary public goods contributions games, and contributions typically level off before 10 periods are reached. Below we reproduce data from Bó (2005) on a voluntary public goods contribution game. As can be seen contributions decline substantially over the 10 rounds, but there is no decline after the seventh round. Hence we will take it that future declines are not a matter of concern.



The second issue is, who to identify as opportunists? The data clearly shows that some members who are willing to make voluntary contributions in early periods reduce their willingness to contribute in later periods. One theory of why this might be the case is that of reciprocal altruism: people are willing to do their share provided others do so, but as they observe others failing to do their share they stop contributing. While this theory has been widely discussed (see, e.g., Fischbacher & Gachter, 2010) as an explanation for declining voluntary contributions over time, it is problematic in the current context because it ought to apply as well to willingness to bear the cost of punishment, and despite the fact there are free riders failing to bear their share of the punishment cost, we do not observe declining willingness to punish over time. Another theory that might account for both facts is a kind of threshold altruism: some acolytes are willing to do their share if the social benefits are great, but do not bother if they are small. Hence they are willing to accept free-riding in punishment because the social benefits are great, but not in contributions because the benefits are small.

[10]Average contributions for both no-punishment and punishment are taken from their Table 3.

A dynamic theory of two types of acolytes, some of whom are willing to "do their bit" regardless of circumstances and some of whom are willing "to do their bit" only if they feel circumstances warrant it is beyond the scope of this paper.

We will start with a strict interpretation in which we use only the final period of data from the no punishment treatments. This means $\phi = 0.25$ with the corresponding $\gamma = 1.14/\phi = 4.56$ solved from $(1-\phi)\times 0 + \phi\gamma = 1.14$. However, the considerations discussed above leave us somewhat agnostic about both ϕ and γ. In particular there may be "circumstantial" acolytes who stop contributing in the no punishment game but continue to punish in the punishment game. This suggests to us that the Fehr and Gächter (2000) estimate of $\phi = 0.47$ may be the better one.[11] As we show that this makes no difference and as a larger value of ϕ works better in other settings, we subsequently accept $\phi = 0.47$. Second, not only may the no punishment output $x_{np} = 1.9$ be too low due to "circumstantial" acolytes, but in addition x_{np} is not terribly well estimated.[12] In particular the standard error on the estimate of 1.9 is 4.1. As the value $x_{np} = 1.9$ does not do a terrific job of explaining the data, we then ask if a slightly higher value does better and find that $x_{np} = 2.4$, less than a quarter of a standard deviation greater than the point estimate does do so. These values imply $\gamma = 3.07$. Hence our bottom line from the calibration will be $\phi = 0.47$ and $\gamma = 3.07$.[13]

5.4 | Calibration

Given ϕ, γ we can solve the mechanism design problem numerically for each value of π. There are three targets we will try to match: the first two are output $x = 12.3$, and welfare. Welfare is reported as 10% higher than the token utility of 21.1 received in the treatment without punishment (result 8), which is to say 23.3 tokens. The third possible target is the *failure rate*, denoted by R. This is defined by $W = (1-R)(v_0 - cx + 3vx) - 10R$, where the factor of 10 is there because each punishment point which costs one token buys only a 10% increase in failure. Using $x = 12.3$ and $W = 23.3$ gives $R = 0.11$.

As utility measured in tokens is not especially interesting, we normalize utility so that it is zero when no public good is produced and one at the maximum possible utility of 32 when everyone donates 20 tokens and there is no punishment. That is, if U is utility in tokens we report welfare $(U-20)/12$. In these units welfare from no punishment of 21.1 tokens becomes 0.10 and from punishment of 23.3 tokens is 0.28 respectively. In other words, the mechanism observed in the data is successful in the sense that it yields 0.18 more utility than that without punishment.

Below we report the results of our calibration

[11]This estimate is broadly consistent with the literature, see Fischbacher and Gachter (2010) or Andreozzi et al. (2020), for example.

[12]Note however that $x_{np} = 1.9$ is large enough that if π were small it would be possible for acolytes to implement the first best $y = 20$.

[13]The units of γ are unclear both theoretically and empirically. It would appear from a theoretical point of view that it would be best measured as payment per unit of time, not payment per match. From information provided in Fehr and Gächter (2000) it appears that a token was worth about \$0.50 US at the time of the experiment and that each match took less than 6 min. This would imply in terms of dollars per hour a γ of about \$15.

Data	ϕ	x_{np}	π	γ	p/γ	x	Welfare	R
FG: 10						12.3	0.28	0.11
	0.25	1.9	0.28	4.56	1.0	12.3	0.39	0.07
	0.47	1.9	0.28	2.43	1.0	12.3	0.39	0.07
	0.47	*2.4*	*0.38*	*3.07*	*1.0*	*12.2*	*0.28*	*0.11*

The first row is the data. The second row is our baseline of $\phi = 0.25, x_{np} = 1.9$ with π chosen to match output x. As can be seen this implies a much lower failure rate than is seen in the data and as a result implies a higher value of welfare. We conclude that this does not look like an optimal mechanism.

The third row simply verifies that it makes no difference whether $\phi = 0.25$ or $\phi = 0.47$.

Finally, in the fourth we search over values of x_{np} and find that with $x_{np} = 2.4$ we can match all three values output, welfare, and the failure rate quite well. The corresponding value of $\pi = 0.38$. This indicates a substantial amount of bad behavior, although as we will present evidence that 10 periods is not enough to find the optimal mechanism in other experiments, it may be that some of this is also due to some confusion over the social norm.

6 | FAIRNESS AND THE EQUAL SPLIT

In this section and the next we examine two games that have been heavily studied in the experimental laboratory: dictator and ultimatum. In both of these two-player games the first mover receives an endowment X and from it offers an amount x to the second mover. In dictator the decision of the first mover is final; in ultimatum the second mover has the option to reject the offer in which case both get zero. We denote by c^i the amount received by each player: $c^1 = X - x$, $c^2 = x$ in dictator or if there is agreement in ultimatum, or zero if the offer is rejected in ultimatum. For both games offers greater than 0 are common, and a $50 - 50$ split is often observed.

What mechanism design problem would result in a 50–50 sharing rule in a dictator or ultimatum game? The answer is that there are several, and indeed we know from the work of Townsend (1994) and Prescott and Townsend (1984) that mechanism design with ex ante uncertainty about types creates a strong tendency towards equal division. Here we highlight two forces working towards equal sharing.

Risk and insurance: Laboratory participants are known to be risk averse over laboratory stakes. If they are ex ante identical then it is socially optimal to share unanticipated gains. In particular, in a dictator game if both participants have an identical risk averse utility function $u(c^i)$ then welfare is $u(X - x) + u(x)$ which is maximized when $x = X/2$.

Incentives and commitment: We know that giving is sensitive to effort (Kahneman et al., 1986). Indeed, even in dictator effort is involved for both parties: the effort in showing up to the laboratory and remaining even when it is discovered that the participant has been assigned to the role of recipient. When there is joint production and effort is complementary, if all the output accrues to one partner there is a commitment problem: ex ante there should be commitment to sharing to provide the partner with incentives to provide effort, but *ex post* the partner who receives the output would prefer to keep it. Social mechanisms can provide the missing commitment. As a simple illustration, suppose there is a joint production function in

which output is $y = V(x^1x^2)^\alpha$ with $\alpha < 1/2$. If h_i is the output share of individual i then individual expected utility is $h_i\, y - x^i$. Fixing the output shares the optimal individual output is shown in Appendix A to be $x^i = (\alpha V)^{1/(1-2\alpha)}(1-h_i)^{\alpha/(1-2\alpha)}h_i^{(1-\alpha)/(1-2\alpha)}$. The social objective function is then

$$V(x^1x^2)^\alpha - x^1 - x^2 = ((1-h_1)h_1)^{\alpha/(1-2\alpha)}(\alpha V)^{2\alpha/(1-2\alpha)}V(1-\alpha).$$

This has a maximum at $h_1 = 1/2$: that is the optimal incentives are provided by an equal sharing rule.

6.1 | Demand for fairness

The simplest and cleanest model is that of risk. To do a quantitative analysis we need a utility function. Here we take the calibration from Fudenberg and Levine (2011): if c denotes laboratory earnings they suggest that a utility function of the form $1-(1+c/C)^{1-\rho}$ with $C =$ \$40 fits the data reasonably well. There is considerable heterogeneity in risk aversion (which we will ignore) and they find that the median coefficient of relative risk aversion ρ is about 1.43. This can be thought of as a measure of the demand for fairness: the greater is ρ the greater the social gain from equalizing income. In dictator, as we shall see, the value of ρ matters little as long as it is positive. By contrast, in ultimatum ρ plays a key role - and $\rho = 1.43$ is not nearly large enough to explain observed behavior through the social mechanism theory outlined above: it predicts considerably more selfish behavior than we observe. Since, as we have indicated, there are additional forces creating demand for fairness we do not view this as an important shortcoming. To account for these additional forces we propose to keep the simple clean risk aversion model but for social utility use the CES utility function with $\rho = \rho_r + \rho_f$ where $\rho_r = 1.43$ and ρ_f is a calibrated additional demand for fairness.

6.2 | Dictator

Dictator games are relatively easy. There is no possibility of punishment: with the standard $X =$ \$10 the theory says that the acolytes should contribute the minimum of \$5 and γ. In Engel (2011)'s meta-study of dictator games "dictators on average give 28.35% of the pie" but for students (the subject population for the public goods and ultimatum experiments we discuss) the meta-regression at the beginning of section 4.6 gives a value of 24.7%. This is remarkably close to 47% of acolytes each giving 50%, which is to say that if $\gamma \geq$ \$5.00 the theory predicts what we see in dictator games.

It is worth pointing out that the theory contends equally well with experiments in which there is an additional option to "take" \$5.00 from the second mover. In this case the free riders should indeed take, while the acolytes offers would be $-5.00 + \gamma$ or \$5.00 if the latter is smaller. Indeed, we can use the results of "take" experiments to get an estimate of γ. In List (2007) adding the "take" option resulted in a drop from giving away \$2.48 to taking of \$1.33 that is a drop of $\$2.48 + \$1.33 = \$3.81$. If we let λ represent the drop in acolytes offers we have $3.81 = \phi\lambda + (1-\phi)5$ so that $\lambda = (\$3.81 - \$2.65)/0.47 = \$2.47$ which says that acolytes

donations drop from \$5 to \$(5 − 2.47) = \$2.53; since the donation is $2.53 = \min\{\gamma - 5, 5\}$ this in turn implies a value of $\gamma = \$5.00 + \$2.53 = \$7.53$.[14]

6.3 | Ultimatum

We review the mechanism design problem, assuming that the endowment X is 10. Individual utility is given by $u(c) = 1 - (1 + c/C)^{1-\rho_r}$ and social utility by $w(c) = 1 - (1 + c/C)^{1-\rho_r-\rho_f}$ where $C = 40$ and $\rho_r = 1.43$. In our reporting we will continue to normalize social utility as a fraction of the maximum, that is, we will multiply by $1/w(5)$ (5 being the equal split of $X = 10$). We denote by q the probability an acolyte rejects an offer on a bad signal, and continue to denote by π the error rate in the signal process. As in our dictator data we discovered γ considerably above \$5 we assume that acolytes are willing to reject any offer on a bad signal (so q can be as high as 1) and are willing to offer \$5 (which is Y in this case) which for efficiency reasons they should do. Hence the objective function is

$$(1 - q\phi\pi)[\phi w(5) + (1 - \phi)(1/2)(w(y) + w(10 - y))]$$

and should be maximized with respect to q, y subject to the constraints that $q \leq 1$ and that for opportunists the utility of conforming to the social norm and offering y is better than deviating and offering zero: $(1 - q\phi\pi)u(10 - y) \geq (1 - q\phi)u(10)$. Since this must hold with equality at the optimum we can compute

$$q = \frac{u(10) - u(10 - y)}{\phi u(10) - \phi\pi u(10 - y)}.$$

6.4 | Calibration

We will now engage in a calibration exercise based on two sources of data. Initially we use data from Roth et al. (1991) as in an earlier behavioral calibration exercise by Levine (1986). We use pooled tenth round data. This data has the advantage that it is described in some detail and has been extensively analyzed in the literature. It also has the property that it does not resemble an optimal mechanism. Is this a failure of the theory or did the participants simply not have adequate time to learn how to implement an optimal mechanism? To assess this we shall use a much longer data series of 40 periods from Duffy and Feltovich (1999).

In the public goods game we had $\phi = 0.47$. In Roth et al. (1991) there is a large jump in the number of offers from \$4.25 to \$4.00 and the fraction of offers \$4.25 or less is 48%. This is

[14]To compare this value of γ to the public goods experiment we need to know how long it took to play and whether overhead time such as instruction time should be included. In the case of the public goods experiment the overhead time was spread over twenty matches. In the dictator experiment there was only one match. If the match including overhead time in dictator took about half an hour this value of γ would be about the same as we calibrated in the public goods experiment. However, it seems unlikely that it would take that long, so it might be that γ is substantially larger here than in the public goods experiment. On the other hand, perhaps in dictator the earlier half hour earnings task should be included. In that case γ might be smaller in dictator than public goods. The difficulty in comparing γ across experiments is a weak point of our calibration.

consistent with the idea that generally acolytes make offers close to \$5.00, so we shall continue to take $\phi = 0.47$.

Our next step is to calibrate ρ_f. We do so by assuming that in Roth et al. (1991) break-even is achieved near the end of 10 rounds as it is in Fehr and Gächter (2000). The smallest value of ρ_f for which break-even is achieved in the tenth round is $\rho_f = 8.27$, considerably larger than the coefficient of relative risk aversion. For our basic calibration we will take a slightly higher value $\rho_f = 8.57$ (which results in $\rho_r + \rho_f = 10$).

As we did in the public goods experiment we can now compute the optimal mechanism as a function of π. As before we target output x, welfare, and a measure of the failure rate $R = q\phi$ which it the probability of rejection conditional on a bad signal.[15] Below are the results of our calibration

Data/case	ρ_f	π	q	x	Welfare	R
RET: 10				4.07	0.83	0.34
	8.57	0.21	0.72	4.07	0.91	0.34
DF out of sample	8.57	0.38	0.47	3.25	0.85	0.22
DF: 11-40				3.63	0.88	0.16

The first row is the data from Roth et al. (1991). The third line is our baseline calibration. Here we choose π from the public good experiment above, $\pi = 0.38$. As can be seen this does not work: output is lower than in the data and the rejection rate much lower. Somewhat surprisingly this results in only a modest increase in welfare. In the second row we choose π so that output is matched. Here the rejection rate is matched but welfare is much too high.

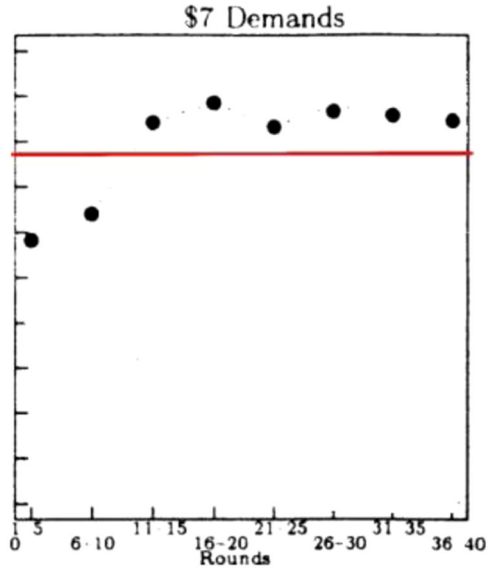
Our conclusion is that the Roth et al. (1991) does not look like an optimal mechanism. Is this because the theory is wrong, or because 10 periods is not long enough to find the mechanism?[16] To answer this we took data from Duffy and Feltovich (1999) for periods 11 through 40. This is reported in the final row of the table.[17] A crucial fact is that output fell and the failure rate dropped by a large amount after the tenth period. With respect to Duffy and Feltovich (1999) we have a pure out of sample forecast. The parameter $\phi = 0.47$ is from both the Fehr and Gächter (2000) and Roth et al. (1991), $\rho_r = 1.43$ is from Fudenberg and Levine (2011), $\rho_f = 8.57$ is from Roth et al. (1991) and $\pi = 0.38$ is from Fehr and Gächter (2000).

[15]The conditional rejection rate in the data is taken to be the rejection rate for "clearly bad signals." We define this as an offer of \$3.25 or less. This was based on inspection of the rejection rates in the Roth et al. (1991) data: there is a clear break at \$3.25. If we take the cutoff at \$3.00 then $R = 0.24$ which while it fits the theory better is implausible in light of the higher rejection rate when the cutoff is \$3.25.

[16]Note that Roth et al. (1991) indicate that their examination of the data leads them to believe that convergence has not yet taken place.

[17]Duffy and Feltovich (1999) do not provide the frequency of offers so we cannot directly compute welfare as we do for Roth et al. (1991). Instead we approximate it using the assumption that ϕ offers of \$5.00 are made and accepted (all such offers were), and that the remaining offers are identical and calibrated to match the mean offer $x = 3.63$ and that these identical offers are rejected at the conditional rejection rate $R = 0.16$. The same procedure applied to the RET10 data with mean offer $x = 4.07$ and $R = 0.34$ yields a welfare value of 0.81 quite close to the value computed from the detailed data of 0.83.

These parameters do a credible job of explaining the Duffy and Feltovich (1999) data although output in the data is somewhat larger than the theory predicts and the failure rate somewhat higher. To give a sense of the failure rate, we reproduce below the acceptance rates for \$3 offers (\$7 demands) as reported in Duffy and Feltovich (1999) together with the red line which is the implied rate from $R = 0.22$ which is to say 0.78 for acceptance.



As in the Roth et al. (1991) data the acceptance rate for the first 10 periods is much lower than required by the theory, but beginning in the 11th period it jumps up substantially to roughly the level required by the theory.

Finally, we ask what happens if we do not constrain π to be the same as in the public good experiment. The second "recalibrated" row of the table chooses π to closely match output in the Duffy and Feltovich (1999) data. This results in a value of, $\pi = 0.21$ lower than we found in the public goods data and roughly the same failure rate. An important observation is this: in Duffy and Feltovich (1999) very little changes in the final thirty rounds. If π represents uncertainty about the social norm we would expect this uncertainty to resolve over thirty periods of play and we see no such thing. Hence again we interpret π as due to bad behavior. This is less clear cut in the public good experiments where play ended in 10 rounds, so it is plausible that π might be smaller here than in the public goods case.

6.5 | Demand for fairness revisited

We calibrated the demand for fairness ρ_f so that the break-even is achieved in Roth et al. (1991). Even with that assumption we do not find evidence that participants are succeeding in solving a mechanism design problem. By contrast with the longer data series in Duffy and Feltovich (1999) we find evidence that after the tenth period they are. This suggests that there might be little reason to calibrate the demand for fairness to the Roth et al. (1991) data. We explored a couple of alternatives for ρ_f, and as we report here the results do not change

substantially. The table below considers alternative values of ρ_f, targeting the Duffy and Feltovich (1999) data reproduced in the final row.[18]

Data/case	ρ_f	π	x	Welfare	R
match	11.57	0.42	3.20	0.83	0.22
	11.57	0.38	3.38	0.85	0.25
DF out of sample	*8.57*	*0.38*	*3.25*	*0.85*	*0.22*
	5.57	0.38	2.91	0.86	0.14
match	5.57	0.31	3.33	0.88	0.22
DF: 11-40			3.63	0.88	0.16

To isolate the effect of ρ_f, in the second, third and fourth rows we hold fixed π. The third row with $\rho_f = 8.57$ is reproduced from table above for ease of comparison. As can be seen raising ρ_f raises both output and the rejection rate, while lowering it lowers both. There is little effect on welfare. If we adjust π to match the rejection rate for the out of sample calibration (rows marked with "match") we see that the higher value of ρ_f does slightly worse on output and welfare and the lower value slightly better, but the effect is modest. Over all it appears that the calibration is relatively robust to substantial changes in ρ_f. Note, however, that even in the Duffy and Feltovich (1999) data risk aversion alone is not enough to explain the data in the strong sense that the break-even point is $\rho_f = 3.37$, still positive.

7 | CONNECTING THE PRIMARY AND THE SECONDARY: CROSS CULTURAL ULTIMATUM

A key implication of our theory is Theorem 3 relating the solution of the primary to the secondary. To assess this using laboratory data we need to compare the same experiment across different societies that have different primaries and hence different levels of internalization ϕ^*. This is delicate because we do not have much information about the primaries: the mechanism design problems may be different, the importance of public goods V_1 may be different, and the monitoring technologies may be different. While there have been many cross-cultural studies there is often little variation in the outcomes and it is hard to assess how the primaries differ. To give a sense of the difficulties, consider the Roth et al. (1991) study from which we analyzed the US data. This study was also conducted in three other societies: Israel, Japan, and Yugoslavia. Despite the fact in some cases the subject populations were college students and in other cases soldiers, and despite the fact that there is variation in output and rejection rates, Fudenberg and Levine (1997) show that from the perspective of participant losses there is not much difference in the outcomes. On top of this we have very little idea how to compare the primaries for these four countries.

As an alternative, we looked at data from the broadest cross-cultural study we know of, namely the the famous study of Henrich et al. (2001) of fifteen small and very different societies. These are

[18]Welfare in the data depends on ρ_f but in the range 5.57 to 11.57 it is 0.88 to two significant digits.

primitive societies and the authors assess an ethnographic variable indicating the gains to cooperation in each society. This is very close to V_1 and because there is so much variation in this variable, we may hope it swamps other relevant differences such as monitoring technology. It is also the case that one of the games they studied was the ultimatum bargaining game, and indeed this was the only game that was used at all sites. Using these data we test the main prediction of the model concerning the extent of punishment in the secondary: namely that it should first increase then decrease as the value of the public good goes up in the primary.

7.1 | Theory of first period play

Unfortunately in the Henrich et al. (2001) study the participants did not play 10 times, let alone thirty times—they played only once. We already know that without substantial experience participants struggle to implement an optimal mechanism. We argue, however, that acolytes are striving to achieve an optimal mechanism even in a single period of play. As we saw before the optimal rejection rate is much lower than is observed in the data: the calibrated conditional rejection rate is 0.22 while the first period conditional rejection rate from Duffy and Feltovich (1999)[19] is 0.46. In early periods acolytes over punish. This makes sense: understanding the necessity of punishment to control the opportunists, to be "safe," acolytes punish sufficiently hard that they know they can get the opportunists under control. This is costly—which is why it is not an optimal mechanism—but given more periods of experience they learn to reduce the punishment: this is what the data shows in both the public goods game and in ultimatum bargaining. This suggests that even in the first period the effect of ϕ should be felt.

We propose a simple model of first-period play. Suppose that q_2^* is optimal with respect to the long-run error rate $\pi_2 = 0.38$. We hypothesize that the observed $\tilde{q}_2$ is proportional to the optimum, that is, $\tilde{q}_2 = \theta q_2^*$. That is, acolytes over punish relative to the optimal mechanism, but their over-punishment reflects the extent to which punishment is desirable in the sense of being part of the optimal mechanism.

Below we give data and theory for our basic calibrated parameters with $\phi = 0.47$. The first row is first period data from Duffy and Feltovich (1999) and the second is the optimal mechanism. We will explain the final column and two rows shortly.

Data/case	q_2	x_2	R	R^u
DF/RET: 1		4.10	0.46	0.26
DF out of sample	0.47	3.25	0.22	0.08
$\theta = 2.09$	0.98		0.46	0.17
$\theta = 2.09, \tilde{\pi}_2$	0.98	3.91	0.46	0.26

We continue with $\rho_f = 8.57$ and $\pi_2 = 0.38$. From $R = \tilde{q}_2\phi$, $R = 0.46$ and $\phi = 0.47$ we get $\tilde{q}_2 = 0.98$, that is the over-punishment factor is $\theta \equiv \tilde{q}_2/q_2^* = 2.09$. The penultimate row of the table shows that this then matches the conditional rejection rate in the data.

[19]First period figures for Duffy and Feltovich (1999) are extrapolated from the averages in the first five and second five rounds.

When ϕ is small the conditional rejection rate $R = \tilde{q}_2\phi$ cannot easily be observed for low values of ϕ since there are very few acolytes, hence very few observations on how frequently they reject. Hence in our cross-disciplinary work we will focus on the unconditional rejection rate $R^u = \tilde{q}_2\pi_2\phi$ which is directly observed. This is reported both for the data and theory in the final table column above.[20] As can be seen the unconditional rejection rate in the data is much higher than for the θ adjusted theory. This leads to our second adjustment. We assume that due to confusion over the social norm the actual error rate $\tilde{\pi}_2$ is larger than the long-term error rate $\pi_2 = 0.38$. As the ratio of the unconditional rejection rate in the data 0.26 to that in the theory 0.17 is $\lambda = 1.53$, we should take $\tilde{\pi}_2 = \lambda\pi_2 = 0.58$.

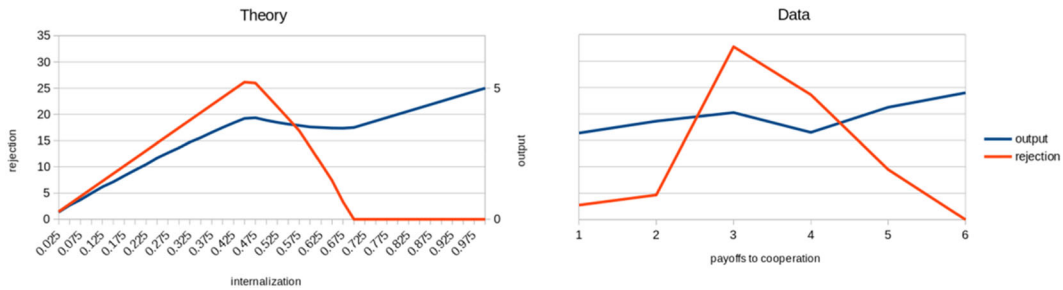
Finally, we turn to the behavior of the opportunists as reflected by the incentive compatible quota y_2. It does not make sense to assume that this is chosen with respect to some long term goal of the acolytes, so we assume that it is chosen optimally with respect to the parameters that exist in the first period, namely $\tilde{q}_2, \tilde{\pi}_2$. This is computed by inverting the incentive constraint

$$\tilde{q}_2 = \frac{u(10) - u(10 - y_2)}{\phi u(10) - \phi\tilde{\pi}_2 u(10 - y_2)}.$$

The final row of the table above carries out the calculation, where it can be seen that output $\tilde{x}_2 = 3.91$ is similar to the value 4.10 observed in the data.

7.2 | Out of sample analysis of cross-cultural ultimatum

With this theory of first period play we now report how output $\tilde{x}_2$ and the unconditional rejection rate $R^u = \tilde{q}_2\pi_2\phi$ vary with internalization ϕ. For ease of reading the unconditional rejection rate R^u is reported in percent. This is plotted side by side with the data from Henrich et al. (2001) (using the same vertical scale) which we discuss subsequently.[21]



[20]The data is taken from round one of Roth et al. (1991) as it is not available in Duffy and Feltovich (1999). If we apply the same approximation procedure used to assess welfare in Duffy and Feltovich (1999) to assess the unconditional rejection rate we get $R^u = 0.24$, quite close to what is observed in the Roth et al. (1991) first period data.
[21]We omit data from one group, the Lamalera because deception was used. The data is taken from Henrich et al. (2004) as it is more conveniently presented there.

The features of the graph are qualitatively like those of Theorem 3. The unconditional rejection rate is not monotone: it initially increases then declines as with many acolytes there is little reason to pay the cost of punishing the few opportunists. Second, output initially rises quite rapidly. Unlike the optimal mechanism for which it flattens out, the over-punishment mechanism has output decline slightly after the initial rise, then rise again. As can be seen, this is driven by the cost of punishment: after the unconditional rejection rate peaks and starts declining output does not rise much, rather the increased internalization is used to reduce punishment costs. Once this is exhausted because punishment is no longer used (recall that $y_2^* = 0$ for high ϕ^* from Theorem 3), output begins again to rise more rapidly.

How does the data in Henrich et al. (2005) compare? The horizontal axis in the figure "payoffs to cooperation" is a categorical variable taking on the values $\mathscr{C} \in \{1, 2, ..., 6\}$. This is an ethnographic variable based on the extent to which each society is judged to benefit from cooperation—or to say the same thing—the importance of public goods in each society. It is conceptually the same as the primary value V_1. The vertical axis is as in the theory, which we have put side-by-side for comparison. As indicated we are making the heroic assumption that the only difference between these different societies is in fact the primary value of V_1. This determines internalization ϕ as an increasing function from Theorem 2. Hence the horizontal axis may also be taken to measure ϕ. There is some anecdotal evidence to support this: according to Henrich et al. (2005) in Ache, the group with the highest value of V_1, "Successful hunters often leave their prey outside the camp to be discovered by others, carefully avoiding any hint of boastfulness." This sounds like a high value of ϕ. Note, however, that the relationship between the horizontal axes for the data and theory is monotone, but there is no reason that it should be linear: we reason to think that V_1 is linear in the ethnographic variable, and the relationship between V_1 and ϕ is not linear.

As can be seen output is qualitatively similar in both figures, initially rising, then declining slightly then rising again. The rejection rate is an inverted U in both cases. Although broadly speaking the figures are quantitatively similar, there are three quantitative anomalies.

The most serious anomaly is that on the left output is much higher than predicted by the theory for the given rejection rate. Specifically, in the data, the first two data points have an average rejection rate of 3.7% and output of 3.50. According to the theory when the rejection rate on the left is 3.7% output should be 0.64, much less than 3.50. We discuss this in the next section. There are also two anomalies which we view as less serious. On the right in the data output begins to rise before the rejection rate reaches zero although according to the theory this should happen only when the rejection rate drops to zero. Finally, the peak of the theoretical rejection rate is somewhat lower than the peak in the data.

7.3 | In sample analysis of cross country ultimatum

The quantitative out of sample analysis fails badly for the two low-V_1 societies $\mathscr{C} \in \{1, 2\}$. Our observation that it is hard to reconcile small rejection rates with relatively high output is not new and has been discussed, for example, in Henrich et al. (2004). One explanation that is discussed is that of risk aversion. One of the anomalous groups were the Orma, studied by Ensminger (2004). Interviews with informants were revealing: "people were obsessed with the possibility that their offer might be refused, in spite of the fact that they thought (correctly) that it was unlikely that people would refuse even a small offer." In particular fairness does not seem to be the key issue.

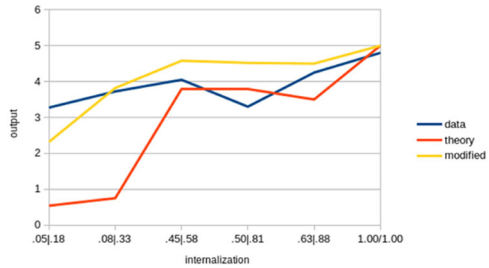
Could it be that risk aversion may be considerably higher in these small societies than the $\rho_r = 1.43$ we used based on studies with college students? This would lead to higher output than predicted by the theory. One problem is that in one of the anomalous societies, the Mapuche, there is independent evidence of risk aversion from Henrich and McElreath (2002). They find that individuals are primarily risk loving rather than risk averse. However, the Mapuche have a higher rejection rate than others with $\mathscr{C} = 2$ and so for them the anomaly is smaller. Without more detailed data it is hard to say more.

Instead we turn to the model. Recall our assumption that $\widetilde{\pi}_2$ has two components, a long run component $\pi_2 = 0.38$ due to "bad behavior" and a "confusion factor" $\lambda = 1.53$. We held these fixed and independent of the society. Does this make sense?.

First, we might consider "bad behavior" among young, rich and bored western college students playing for relatively low stakes versus older, poorer participants some of whom traveled a considerable distance to participate and who played for relatively high stakes. We would surely expect more "bad behavior" in the former. This suggests that in the cross-cultural experiments π_2 might be considerably lower than for the western data we calibrated to.

Second, does it make sense that the confusion over the social norm as measured by λ is a constant? Suppose output is x_2. Participants probably have a rough idea what this number is and that the social norm y_2 lies to the left. But for larger x_2 there is greater scope for confusion: if $x_2 = 4.75$ then there are a broad range of $y_2's$ lying to the left, while if $x_2 = 2.5$ there are considerably fewer possible social norms. This suggests that we might have $\lambda(x_2)$ which is increasing in x_2.

To examine the potential impact of these two factors we chose $\pi_2 = .15$ (to match output for $\mathscr{C} = 2$) and assumed that for $x_2 < 4$ we have no confusion, $\widetilde{\pi}_2 = \pi_2$ while for $x_2 > 4$ the noise $\widetilde{\pi}_2$ remains as it is in our original calibration.[22] The comparison of output between the theory and the data is reported below.



Here ϕ is estimated by matching the unconditional rejection rate in the data to that in the theory. The blue line is output in the data, the red line in the out of sample theory and the yellow line in the modified theory as just described. The labels on the horizontal axis correspond to the out of sample theory and the modified theory respectively. Hence, for example, for $\mathscr{C} = 2$ the unconditional rejection rate in the data is $R^u = 4.7\%$. In the out of sample theory this rejection rate to the left of the peak occurs when $\phi = 0.08$, and for the in sample theory when $\phi = 0.33$. The height of the blue curve matches the mean output for $\mathscr{C} = 2$ in the data $x_2 = 3.73$, the height of the red curve is the output computed from the out of sample theory for $\phi = 0.08$,

[22]Specifically we take $\lambda = 3.88$. This is the factor needed to reconcile π_2 which the observed unconditional rejection rates in the western data. Alternatively this can be thought of as calibrating λ to the unconditional rejection rate for the intermediate society $\mathscr{C} = 4$ which is similar to that in the western data.

that is $x_2 = 0.75$, which as we know is far too low, and the height of the yellow curve is computed from the in sample theory for $\phi = 0.33$ which gives $x_2 = 3.82$ and closely matches the data because we calibrated π_2 so this would be the case.

Quantitatively the modified theory does much better than the out of sample theory, although output initially rises too fast then remains too high. For the two low V_1 societies the modified theory does much better than the out of sample theory, which is not surprising since it was designed for this purpose. For the four high V_1 societies the modified theory does clearly worse for $\mathscr{C} = 4$ although neither theory really captures the dip in the data, it does slightly worse for $\mathscr{C} = 3$ and considerably better for $\mathscr{C} < 3$. Hence we take both theories as reasonable for $\mathscr{C} \in \{3, 4, 5, 6\}$ while the modified theory does a good job for $\mathscr{C} \in \{1, 2\}$ as well.

Our interpretation of the data seen through lens of social mechanisms with internalization is rather different than that taken by Henrich et al. (2005). Their view is that greater objective incentive for cooperation outside the laboratory leads to greater fairness inside the laboratory. This does not predict the lack of monotonicity of offers and punishment that our theory predicts and that we observe in the data. While 14 observations of widely differing societies and a handful of ultimatum games played in each society under difficult conditions cannot be persuasive, the theory of social mechanisms provides a much more detailed and sharper account of what to look for in the data. An interesting case is that of the Ache. This society has the highest V_1, the highest x_2 and the unconditional rejection rate is at the minimum $R^u = 0$. It is also described as a highly homicidal society in which in-group homicide is a common cause of death. Our interpretation is that this corresponds to a high cost of punishment in the primary. This is exactly as our theory indicates: while punishment in the secondary eventually declines with V_1 punishment in the primary strictly increases.

8 | CONCLUSION

We conclude by indicating how the ideas in this paper fit into the broader literature of experimental and behavioral economics. Writers such as Gintis et al. (2003) and Roemer (2015) point to evolutionary reasons why punishment might be "hard-wired." Experimentalists such as Fehr and Gächter (2000) similarly argue that intrinsic preferences for reciprocal altruism "do unto others as they have done unto you" are observed in the laboratory. We do not doubt that small children do not need to be taught to punish the theft of a toy. Never-the-less social norms must - and do - specify punishment levels scaled to the nature of the offense, the benefit of deviating, and the chances of getting caught. Hence our approach of treating the choice of punishment as the solution to a mechanism design problem. In particular in our setting acolytes carry out punishments because they are useful in solving the social problem of public goods provision, not because of an intrinsic desire for revenge.

We examine a particularly simple stark theory of internalization based on warm glow giving and study the trade-off between the use of incentives and internalization. We show that the idea that in the laboratory participants solve mechanism design problems subject to uncertainty but making good use of internalization is consistent with what we see. In particular we find that internalization is important in alleviating the need to provide incentives to monitors and that data from laboratory experiments is broadly consistent with a fraction of the population internalizing the social norm. Moreover, our theory predicts that as the value of the public good increases, incentives and internalization are complements in the society at large but at a threshold value of the public good they become substitutes in the laboratory. This supports the

prediction that the extent of punishment in the laboratory is not monotone in the "value of cooperation" in the society at large.

ACKNOWLEDGMENTS

We would like to thank Luciano Andreozzi, Marco Casari, Andrea Ichino, Andrea Mattozzi, Rohini Somanathan and seminar audiences at WUSTL, Warwick, Queen's, the Paris Institutions Conference, the Zurich Political Economy Conference, Delhi School of Economics and the University of Trento. We gratefully acknowledge support from the EUI Research Council and MIUR PRIN 20103S5RN3.

DATA AVAILABILITY STATEMENT

The data is from published sources reported in the article.

REFERENCES

Abreu, D., & Rubinstein, A. (1988). The structure of Nash equilibrium in repeated games with finite automata. *Econometrica*, 1259–1281.

Akerlof, G. A., & Kranton, R. E. (2000). Economics and identity. *The Quarterly Journal of Economics*, *115*(3), 715–753.

Andreoni, J. (1990). Impure altruism and donations to public goods: A theory of warm-glow giving. *Economic Journal 100*: 464–477.

Andreozzi, L., Faillo, M., & Saral, A. S. (2020). On altruism, reciprocal and not. A dictator game experiment. Mimeo Trento.

Belloc, M., Drago, F., & Galbiati, R. (2016). Earthquakes, religion, and transition to self-government in Italian cities. *The Quarterly Journal of Economics*, *131*, 1875–1926.

Bénabou, R., & Tirole, J. (2006). Incentives and prosocial behavior. *The American Economic Review*, *96*(5), 1652–1678.

Bigoni, M., Bortolotti, S., Casari, M., Gambetta, D., & Pancotto, F. (2016). Amoral familism, social capital, or trust? The behavioural foundations of the Italian North-South divide. *The Economic Journal*, *126*, 1318–1341.

Block, J. I., & Levine, D. K. (2016). Codes of conduct, private information and repeated games. *International Journal of Game Theory*, *45*, 971–984.

Bó, P. D. (2005). Cooperation under the shadow of the future: experimental evidence from infinitely repeated games. *American Economic Review*, *95*, 1591–1604.

Boldrin, M., Christiano, L. J., & Fisher, J. D. (2001). Habit persistence, asset returns, and the business cycle. *American Economic Review*, 149–166.

Bowles, S., & Gintis, H. (1976). *Schooling in capitalist America* (Vol. 57). Basic Books.

Campbell, J. Y., & Cochrane, J. H. (1999). By force of habit: A consumption-based explanation of aggregate stock market behavior. *Journal of Political Economy*, *107*, 205–251.

Cason, T. N., Sheremeta, R. M., & Zhang, J. (2012). Communication and efficiency in competitive coordination games. *Games and Economic Behavior*, *76*, 26–43.

Coase, R. H. (1960). The problem of social cost. *Journal of Law and Economics*, *3*, 1–44.

Constantinides, G. M. (1990). Habit formation: A resolution of the equity premium puzzle. *Journal of Political Economy*, *98*, 519–543.

Cremer, J., & McLean, R. P. (1988). Full extraction of the surplus in Bayesian and dominant strategy auctions. *Econometrica*, *56*, 1247–1257.

Duffy, J., & Feltovich, N. (1999). Does observation of others affect learning in strategic environments? An experimental study. *International Journal of Game Theory*, *28*, 131–152.

Dutta, R., Levine, D. K., & Modica, S. (2017). *Peer monitoring, ostracism and the internalization of social norms*. EUI.

Engel, C. (2011). Dictator games: A meta study. *Experimental Economics*, *14*, 583–610.

Ensminger, J. (2004). Market integration and fairness: Evidence from ultimatum, dictator, and public goods experiments in East Africa. In J. P. Henrich (Ed.), *Foundations of human sociality: Economic experiments and ethnographic evidence from fifteen small-scale societies* (pp. 356–381). Oxford University Press.

Feddersen, T., & Sandroni, A. (2006). A theory of participation in elections. *American Economic Review*, *96*, 1271–1282.

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

Fehr, E., & Schmidt, K. M. (1999). A theory of fairness, competition and cooperation. *Quarterly Journal of Economics*, *114*, 817–868.

Fischbacher, U., & Gachter, S. (2010). Social preferences, beliefs, and the dynamics of free riding in public goods experiments. *American Economic Review*, *100*, 541–56.

Fudenberg, D., Levine, D., & Maskin, E. (1994). The folk theorem with imperfect public information. *Econometrica*, *62*(5), 997–1039.

Fudenberg, D., & Levine, D. K. (1997). Measuring players' losses in experimental games. *Quarterly Journal of Economics*, *112*, 507–536.

Fudenberg, D., & Levine, D. K. (2011). Risk, delay, and convex self-control costs. *AEJ Micro*, *3*, 34–68.

Fudenberg, D., Levine, D. K., & Pesendorfer, W. (1998). When are non-anonymous players negligible. *Journal of Economic Theory*, *79*, 46–71.

Gale, D., & Sabourian, H. (2005). Complexity and competition. *Econometrica*, *73*, 739–769.

Gintis, H., Bowles, S., Boyd, R., & Fehr, E. (2003). Explaining altruistic behavior in humans. *Evolution and Human Behavior*, *24*, 153–172.

Harsanyi, J. C. (1973). Games with randomly disturbed payoffs: A new rationale for mixed-strategy equilibrium points. *International Journal of Game Theory*, *2*, 1–23.

Henrich, J., & McElreath, R. (2002). Are peasants risk-averse decision makers? *Current Anthropology*, *43*, 172–181.

Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H., & McElreath, R. (2001). In search of homo economicus: Behavioral experiments in 15 small-scale societies. *The American Economic Review*, *91*(2), 73–78.

Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H., & McElreath, R. (2004). Overview and Synthesis. In J. P. Henrich, R. Boyd, S. Bowles, E. Fehr, C. Camerer, & H. Gintis (Eds.), *Foundations of human sociality: Economic experiments and ethnographic evidence from fifteen small-scale societies*. Oxford University Press on Demand.

Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H., & McElreath, R. (2005). Economic man' in cross-cultural perspective: Behavioral experiments in 15 small-scale societies. *Behavioral and Brain Sciences*, *28*, 795–815.

Jackson, M. O. (2010). *Social and economic networks*. Princeton University Press.

Kahneman, D., Knetsch, J., & Thaler, R. (1986). Fairness as a constraint on profit-seeking: Entitlements in the market. *American Economic Review*, *76*, 728–741.

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

Levine, D., & Modica, S. (2016). Peer discipline and incentives within groups. *Journal of Economic Behavior and Organization*, *123*, 19–30

Levine, D., & Modica, S. (2017). Size, Fungibility, and the strength of lobbying organizations. *European Journal of Political Economy*, *49*, 71–83.

Levine, D. K. (1986). Modeling altruism and spitefulness in experiments. *Review of Economic Dynamics*, *1*, 593–622.

Levine, D. K. (2012). *Is behavioral economics doomed?: The ordinary versus the extraordinary*. Open Book Publishers.

Levine, D. K., & Mattozzi, A. (2020). Voter turnout with peer punishment. *American Economic Review*, *110*, 3298–3314.

Levine, D. K., & Palfrey, T. R. (2007). The paradox of voter participation? A laboratory study. *American Political Science Review*, *101*, 143–158.

List, J. A. (2007). On the interpretation of giving in dictator games. *Journal of Political Economy*, *115*, 482–493.

Meyer, C. J., & Tripodi, E. (2017, October 24). *Sorting into incentives for prosocial behavior*. Available at SSRN: https://ssrn.com/abstract=3058195

Muthoo, A. (1996). A bargaining model based on the commitment tactic. *Journal of Economic Theory*, *69*, 134–152.

Olson Jr., M. (1965). *The Logic of collective action: Public goods and the theory of groups*. Harvard Economic Studies.

Ostrom, E. (1990). *Governing the commons: The evolution of institutions for collective action*. Cambridge university press.

Palfrey, T. R., & Prisbrey, J. E. (1997). Anomalous behavior in public goods experiments: How much and why? *American Economic Review*, 829–846.

Ponemon, L. A. (1993). Can ethics be taught in accounting? *Journal of Accounting Education*, *11*, 185–209.

Prescott, E. C., & Townsend, R. M. (1984). Pareto optima and competitive equilibria with adverse selection and moral hazard. *Econometrica*, 21–45.

Rahman, D. (2012). But who will monitor the monitor? *American Economic Review*, *102*(6), 2767–2797.

Rand, D. G., Greene, J. D., & Nowak, M. A. (2012). Spontaneous giving and calculated greed. *Nature*, *489*, 427–430.

Robson, A. J. (1990). Efficiency in evolutionary games: Darwin, Nash and the secret handshake. *Journal of Theoretical Biology*, *144*, 379–396.

Roemer, J. (2015). Kantian optimization: An approach to cooperative behavior. *Journal of Public Economics*, *127*(C), 45–57.

Rogers, V., & Smith, A. (2008). An examination of accounting majors' ethical decisions before and after an ethics course requirement. *Journal of College Teaching and Learning*, *5*, 49–54.

Roth, A. E., Prasnikar, V., Okuno-Fujiwara, M., & Zamir, S. (1991). Bargaining and market behavior in Jerusalem, Ljubljana, Pittsburgh, and Tokyo: An experimental study. *American Economic Review*, 1068–1095.

Schelling, T. C. (1956). An essay on bargaining. *The American Economic Review*, *46*(3), 281–306.

Skarbek, D. (2014). *The social order of the underworld: How prison gangs govern the American penal system*. Oxford University Press.

Stahl, D. O., & Wilson, P. W. (1995). On players' models of other players: Theory and experimental evidence. *Games and Economic Behavior*, *10*, 218–254.

Tangney, J. P., & Dearing, R. L. (2003). *Shame and guilt*. Guilford Press.

Tangney, J. P., Stuewig, J., & Mashek, D. J. (2007). Moral emotions and moral behavior. *Annual Review of Psychology*, *58*, 345–372.

Tirole, J. (2009). Cognition and incomplete contracts. *American Economic Review*, *99*(1), 265–94.

Tisserand, J. C., Cochard, F., & LeGallo, J. (2015). *Altruistic or strategic considerations: A meta-analysis on the ultimatum and dictator games*. CRESE, Université de Franche-Comté.

Townsend, R. M. (1994). Risk and insurance in village India. *Econometrica*, 539–591.

Turner, J. H., & Stets, J. E. (2005): *The sociology of emotions*. Cambridge University Press.

How to cite this article: Dutta, R., Levine, D. K., & Modica, S. (2021). The whip and the Bible: Punishment versus internalization. *J Public Econ Theory*. 23, 858–894. https://doi.org/10.1111/jpet.12540

APPENDIX A

In this Appendix the numbering is that of the text.

Lemma 1. *The solution to* $\max_\theta Vf(a + b\theta) - c\theta$ *subject to* $0 \leq \theta \leq \Theta$ *is unique and given by* $\theta^* = (1/b)(g((1/V)(c/b), a, b\Theta) - a)$. *The function* $g(\mu, \underline{x}, X)$ *is continuous and*

increasing[23] *in* $\underline{x}, X$. *It satisfies* $x \leq g(\mu, \underline{x}, X) \leq x + X$ *and for* $x < g(\mu, \underline{x}, X) < \underline{x} + X$ *it is smooth and strictly decreasing in* μ.

Proof. Taking $x = a + b\theta$ the objective function is $Vf(x) - c(x - a)/b$ while the constraint is $a \leq x \leq a + b\Theta$. If we linearly transform the objective function dividing by V and subtracting $(1/V)ca/b$ we get the equivalent objective function $f(x) - (1/V)(c/b)x$. Hence $x^* = g((1/V)(c/b)), a, b\Theta)$. Substituting into $\theta^* = (x^* - a)/b$ give the desired result.

The properties of g follow directly from the properties of f and the definition of g. □

Theorem 1. *At the optimal solution if* $\phi = 0$ *then* $y^* = 0$ *and* φ,P *do not matter. When* $\phi > 0$ *then* $\varphi^* = (1/(\phi\gamma))g(1/V, 0, \phi\gamma)$ *and*

1. *If* $Vf'(\phi\gamma) \leq 1$ *the optimal solution is first best with* $y^* = 0, P^* = 0$.
2. *If* $Vf'(\phi\gamma) > 1$ *the solution is* second best *with* $\varphi^* = 1$,

$$y^* = g\left(\frac{1+M}{V}, \gamma\phi, \frac{(1-\pi)\phi\gamma}{\psi}\right) - \gamma\phi,$$

and

$$P^* = \frac{y^*}{\phi(1-\pi)}.$$

Moreover, the maximized utility is concave and increasing in ϕ. *Finally,* $y^* \leq (1 - \pi)\phi\gamma/\psi$.

Proof. Consider first the production problem. Observe that the probability of being punished is equal to the probability that the monitor is an acolyte times the probability of a bad signal. Hence for an opportunist the cost of meeting target y is $y + \phi\pi P$, while the best alternative of producing zero costs ϕP, resulting in the incentive constraint $y + \phi\pi P \leq \phi P$ or $y \leq \phi(1 - \pi)P$. As indicated above whenever it is incentive compatible for an opportunist to produce y it is incentive compatible for an acolyte to produce up to $y + \gamma$, that is up to $\varphi = 1$. Therefore, a norm (φ, y, P) with $0 \leq \varphi \leq 1$ is incentive compatible for both types of producers if and only if $y \leq \phi(1 - \pi)P$.

If $\phi = 0$ the only feasible $y = 0$ and $U = 0$ for any P, φ, as in the statement. Now assume $\phi > 0$. Since P should be minimized we get $P = y/[\phi(1 - \pi)]$. Incentive compatibility for monitoring requires $\psi P \leq \gamma$, which inserting the value of P from above reads

$$y \leq (1 - \pi)\phi\gamma/\psi. \tag{A1}$$

This is the final result indicated above.

[23]For brevity increasing and decreasing without qualification always mean weakly so.

Now the monitoring cost of output $y + \phi\varphi\gamma$ is $(1 + \psi)\phi\pi P = My$, so the objective function is

$$U = Vf(y + \phi\varphi\gamma) - (y + \phi\varphi\gamma) - My. \tag{A2}$$

This has to be maximized with respect to y, φ subject to the constraints $y \leq (1 - \pi)\phi\gamma/\psi$ and $0 \leq \varphi \leq 1$.

Since the objective function is concave and the constraint set convex, we see immediately that the maximized objective is concave in ϕ. It is increasing in ϕ: because utility depends only on $x = y + \phi\varphi\gamma$ and y and the feasibility restrictions $x \leq y + \phi\gamma$ and $y \leq (1 - \pi)\phi\gamma/\psi$ are both relaxed as ϕ is increased.

From the objective function we see that φ is a dominant technology over y: that is, increasing output by increasing y has an associated monitoring cost of $(1 + \psi)\pi y/(1 - \pi)$ and φ does not. In particular if at the optimum $\varphi < 1$ then $y = 0$ otherwise output $y + \phi\varphi\gamma$ could be held fixed and utility increased by lowering y and increasing φ. Next we show that $\varphi < 1$ when $Vf'(\phi\gamma) \leq 1$. This occurs because there is also a resource cost of producing output when the designer faces the first best problem of maximizing $Vf(x) - x$. If $Vf'(\phi\gamma) \leq 1$ the solution to this problem is feasible and obtained by taking $y^* = 0$ and from Lemma 1 with $a = 0, b = \phi\gamma, c = \phi\gamma$ choosing φ^* as stated in the proposition.

The solution $\varphi^* = (1/(\phi\gamma))g(1/V, 0, \phi\gamma)$ has the property that $\varphi^* = 1$ for $Vf'(\phi\gamma) \geq 1$. When that is the case it may be optimal to choose $y^* > 0$: we should fix $\varphi^* = 1$ and maximize U in A2 with respect to y under the constraint $y \leq (1 - \pi)\phi\gamma/\psi$. Applying Lemma 1—with $a = \phi\gamma, b = 1, c = [1 + M)]$—the given solution results. From the definition of g if $y^* > 0$ this solution satisfies

$$Vf'(y^* + \phi\gamma) - 1 - M \geq 0$$

so $Vf'(y^* + \phi\varphi\gamma) - 1 > 0$ implying $\varphi^* = (\phi\gamma)^{-1}(g(1/V, y, \phi\gamma) - y^*) = 1 = (\phi\gamma)^{-1}g(1/V, 0, \phi\gamma)$, as it should. □

Corollary 1. *If $Vf'(\phi\gamma) > 1$ then $\varphi^* = 1$ and total output $x^* = y^* + \phi\gamma$ is increasing in V, ϕ and decreasing in π, ψ. Define $\hat{\phi}$ by $Vf'(\chi\hat{\phi}) = 1 + M$. For $\phi < \hat{\phi}$ the optimal quota y^* and punishment P^* are increasing in ϕ and for $\phi > \hat{\phi}$ they are decreasing.*

Proof. From Theorem 1 we know that if $Vf'(\phi\gamma) > 1$ then the solution has $\varphi^* = 1$ and

$$y^* = g\left(\frac{1}{V}\left(1 + (1 + \psi)\frac{\pi}{1 - \pi}\right), \gamma\phi, \frac{(1 - \pi)\phi\gamma}{\psi}\right) - \gamma\phi, \quad P^* = \frac{y^*}{\phi(1 - \pi)}$$

so since total output is $y^* + \phi\gamma$ the first part follows from Lemma 1 and Theorem 1. From the two cited results it also follows that y^* is decreasing in ϕ in the interior (when ϕ is large) but increasing at the upper bound (when ϕ is small). The condition given in the result is the transition between the interior and upper bound.

The final part follows from the fact that from Theorem 1 P^* is increasing in y^*. □

Lemma 2. *The optimal primary social mechanism has* $\varphi^* = 1$.

Proof. Suppose $\varphi^* < 1$. From Theorem 1 #2 the solution must be first best with $y^*, P^* = 0$ and output $x^* = \phi^*\varphi^*\gamma$. If $\phi^* > 0$ we may increase φ and decrease ϕ keeping y^*, P^*, x^* all fixed. Since ϕ has marginal cost H and φ has none this strictly increases the objective function. On the other hand if $\phi^* = 0$ then φ does not matter, so we may as well take it equal to 1. □

Theorem 2. *If* $H < \gamma M$ *then* $\phi^* = (1/\gamma)g((1/V)(1 + H/\gamma), 0, \gamma)$ *and the optimal quota is*

$$y^* = g\left(\frac{1+M}{V}, \gamma, \chi - \gamma\right) - \gamma$$

which is equal to zero if $\phi^* < 1$.

If $H > \gamma M$ *then*

$$\phi^* = \frac{1}{\chi} g\left(\frac{1}{V}\frac{\chi + (1+\psi)\pi\gamma/\psi + H}{\chi}, 0, \chi\right)$$

and $y^* = (1-\pi)\phi^*\gamma/\psi$.

Proof. The partial derivatives of the objective function are

$$\partial W_1/\partial y = Vf'(y + \phi\gamma) - 1 - (1+\psi)\frac{\pi}{1-\pi}$$
$$\partial W/\partial \phi = \gamma(Vf'(y+\phi\gamma) - 1) - H = \gamma\left(\partial W/\partial y + (1+\psi)\frac{\pi}{1-\pi}\right) - H.$$

If follows directly that if $H < \gamma(1+\psi)\pi/(1-\pi) = \gamma M$ then $\partial W/\partial\phi \le 0$ implies $\partial W/\partial y < 0$ hence at the optimum, if $\phi < 1$ so that $\partial W/\partial\phi \le 0$, we have $\partial W/\partial y < 0$ so $y^* = 0$. When $y^* = 0$Lemma 1 gives the expression for ϕ^* in the statement. If $\phi = 1$ (with $\varphi^* = 1$ as well by Lemma 2) we can write the objective function as $W = Vf(y+\gamma) - y - \gamma - y(1+\psi)\pi/(1-\pi) - H$. Applying Lemma 1 gives the expression for y^*. Since $H < \gamma(1+\psi)\pi/(1-\pi) = \gamma M$ if $\phi < 1$ then this expression gives $y^* = 0$ so is valid in both cases. On the other hand $\phi^* = 1$ iff $\gamma Vf'(y + \phi^*\gamma) \ge \gamma + H$ in which case $\gamma Vf'(\phi^*\gamma) \ge \gamma + H$ so $(1/\gamma)g((\gamma+H)/(V\gamma), 0, \gamma) = 1 = (1/\gamma)[g((\gamma+H)/(V\gamma), y, \gamma) - y^*]$.

It similarly follows that if $H > \gamma(1+\psi)\pi/(1-\pi) = \gamma M$ then the constraint $y \le (1-\pi)\phi\gamma/\psi$ binds. Indeed in this case if $\partial W/\partial y \le 0$ then $\partial W/\partial\phi < 0$, so either $\partial W/\partial\phi < 0$ so $\phi^* = 0$ or $\partial W/\partial y > 0$ hence the constraint again binds. This gives the objective function

$$W = Vf((1-\pi)\phi\gamma/\psi + \phi\gamma) - ((1-\pi)\phi\gamma/\psi + \phi\gamma) - (1+\psi)\pi\phi\gamma/\psi - H\phi$$

which making use of $\chi = (1-\pi)\gamma/\psi + \gamma$ is as given above, so the final result follows as well from Lemma 1. □

Corollary 2. *In the primary problem internalization ϕ^*, the production quota y^*, total output $x^* = y^* + \phi^*\gamma$ and punishment P^* are increasing in V. Total output x^* is increasing in γ and decreasing in π. If $H > \gamma M$, punishment P^* is constant in V, and for $0 < \phi^* < 1$ the optimal ϕ^*, y^*, x^* strictly increase.*

Proof. Internalization and the production quota follow directly from Theorem 2, with total output the immediate consequence.

Punishment is given by $P^* = y^*/[\phi^*(1-\pi)]$ from Theorem 1. By Theorem 2 if $H < \gamma M$ then either $y^* = 0$ so $P^* = 0$ or $\phi^* = 1$ in which case y^* is increasing in V so P^* is as well. If $H > \gamma M$ then the constraint binds so $y^* = (1-\pi)\phi^*\gamma/\psi = \phi^*(\chi - \gamma)$ so $P^* = \gamma/\psi$ is independent of V.

To prove the assertion on total output x^* observe from Theorem 2 if $H < \gamma M$ then total output is

$$x^* = g\left(\frac{1}{V}(\gamma + H)/\gamma, 0, \gamma\right) + g\left(\frac{1}{V}(1 + (1+\psi)\pi/(1-\pi)), \gamma, \chi - \gamma\right) - \gamma$$

and if $H > \gamma M$ then total output is

$$x^* = g\left(\frac{1}{V}\frac{\chi + (1+\psi)\pi\gamma/\psi + H}{\chi}, 0, \chi\right).$$

In both cases the assertion follows from the properties of g in Lemma 1. □

Theorem 3. *If $H > \gamma_1 M_1$ and $\underline{V}_1 < V_1 < \overline{V}_1$ then as V_1 increases*

(i) *Acolytes ϕ^*, output x_1^*, and the quota y_1^* all strictly increase, and punishment P_1^* is constant.*

If in addition $V_2 f'(\phi^*\gamma) > 1 + M_2$ there are intermediate cutoffs $\underline{V}_1 < V_1^m < V_1^M < \overline{V}_1$ such that

(ii) *for $\underline{V}_1 < V_1 < V_1^m$ output x_2^* and the quota y_2^* strictly increase while punishment P_2^* is constant. For $V_1^m < V_1 < V_1^M$ output x_2^* is constant and the quota y_2^* and punishment P_2^* strictly decrease. For $V_1^M < V_1 < \overline{V}_1$ output x_2^* strictly increases, the quota $y_2^* = 0$, and punishment is constant at zero.*

Proof. Part (i) is simply a restatement of the relevant portion of Corollary 2.

We know from the primary that as V_1 goes from $\underline{V}_1$ to $\overline{V}_1$ internalization ϕ^* strictly increases from 0 to 1. From Theorem 1 we know that the solution of the secondary is given by

$$y_2^* = g\left(\frac{1}{V_2}(1 + M_2), \gamma_2\phi^*, (\chi_2 - \gamma_2)\phi^*\right) - \gamma_2\phi^*.$$

Recall that $V_2 f'(\chi_2 \hat{\phi}_2) = 1 + M_2$ and define $V_2 f'(\gamma_2 \tilde{\phi}_2) = 1 + M_2$. Note that by our assumptions $0 < \hat{\phi}_2 < \tilde{\phi}_2 < 1$. Hence y_2^* takes on one of three values, with corresponding total output from $x_2^* = y_2^* + \phi^* \gamma_2$ and punishment from $P_2^* = (1/(1 - \pi_2))(y_2^*/\phi^*)$.

For small $\phi^* < \hat{\phi}_2$ (small V_1) it is

$$y_2^* = (\chi_2 - \gamma_2)\phi^*, x_2^* = \chi_2 \phi^*, P_2^* = (1/(1 - \pi_2))(\chi_2 - \gamma_2)$$

for intermediate $\hat{\phi}_2 < \phi^* < \tilde{\phi}_2$ (intermediate V_1) it is

$$y_2^* = [f']^{-1}\left(\frac{1}{V_2}(1 + M_2)\right) - \gamma_2 \phi^*, \quad x_2^* = [f']^{-1}\left(\frac{1}{V_2}(1 + M_2)\right)$$
$$P_2^* = (1/(1 - \pi_2))\left[(1/\phi^*)[f']^{-1}\left(\frac{1}{V_2}(1 + M_2)\right) - \gamma_2\right]$$

while for large $\phi^* > \tilde{\phi}_2$ (large V_1) it is

$$y_2^* = 0, x_2^* = \gamma_2 \phi^*, P_2^* = 0.$$

This gives the desired results. □

Lemma 3. *The expected number of potential punishers conditional on a bad signal is* $Q = 3(1 - \pi) + \pi^2$.

Proof. The second row of the table below lists for a particular participant i who has a bad signal the probability that one of the other three has a bad signal.

Others with bad signals	**0**	**1**	**2**	**3**
Probability	$(1 - \pi)^3$	$3(1 - \pi)^2\pi$	$3(1 - \pi)\pi^2$	π^3
Number punishing	3	2	4/3	1

The final row of the table indicates how many opponents are potentially willing punish i. If i has the only bad signal all three opponents will potentially punish her (total 3). If there is one other bad signal then the two without bad signal each give half a punishment to the two with bad signals, and the one with a bad signal gives a full punishment to i (she does not punish herself), so total in this case is $1/2 + 1/2 + 1$. If there are two other bad signals then the one without a bad signal gives 1/3rd punishment and the two with bad signal each give half a punishment to the other two with bad signals, with total $1/3 + 2 \cdot 1/2 = 4/3$. Finally, if there are three other bad signals then each gives 1/3rd punishment. To compute the expected number of potential punishers, observe that if the numbers in the final row were $3, 2, 1, 0$ the expectation would be $3(1 - \pi)$. Hence the actual expectation is $Q = 3(1 - \pi) + (1/3)3(1 - \pi)\pi^2 + \pi^3 = 3(1 - \pi) + \pi^2$. □