MANAGEMENT SCIENCE
Vol. 56, No. 1, January 2010, pp. 57–70
ISSN 0025-1909 | EISSN 1526-5501 | 10 | 5601 | 0057

informs

DOI 10.1287/mnsc.1090.1081

Another Hidden Cost of Incentives: The Detrimental Effect on Norm Enforcement

Andreas Fuster
Department of Economics, Harvard University, Cambridge, Massachusetts 02138;
and Federal Reserve Bank of Boston, Boston, Massachusetts 02210, afuster@fas.harvard.edu

Stephan Meier
Graduate School of Business, Columbia University, New York, New York 10027;
and Federal Reserve Bank of Boston, Boston, Massachusetts 02210, sm3087@columbia.edu

Monetary incentives, such as subsidies or bonuses, are often considered as a way to foster contributions to public goods in society and firms. This paper investigates experimentally the effect of private contribution incentives in the presence of a norm enforcement mechanism. Norm enforcement through peer punishment has been shown to be effective in raising contributions by itself. We test whether and how (centrally provided) private incentives interact with (decentralized) punishment, both of which affect subjects' monetary payoffs. The results of our experiment show that private incentives for contributors can reduce the effectiveness of the norm enforcement mechanism: Free riders are punished less harshly in the treatment with incentives, and as a consequence, average contributions to the public good are no higher than without incentives. This finding ties to and extends previous research on settings in which monetary incentives may fail to have the desired effect.

Key words: public goods; prosocial behavior; norm enforcement; hidden cost of incentives; experimental economics
History: Received April 6, 2009; accepted August 2, 2009, by Teck-Hua Ho, decision analysis. Published online in *Articles in Advance* October 16, 2009.

1. Introduction

Prosocial behavior, such as making private contributions to public goods, is crucial for the proper functioning of society and the efficiency of organizations. Various activities performed by humans, from hunting to holding a potluck party, require discretionary contributions of the group members to be successful. Many common pool resources become exhausted if individuals do not refrain from consuming the privately optimal, but socially suboptimal, amount. Similarly, the success of organizations depends on members' willingness to take unselfish, efficiency-enhancing actions, because it is often impossible to fully control behavior through contracts. However, the incentives to free ride in such situations generally make it difficult to sustain high levels of prosocial behavior. Recent research shows that peer punishment or norm enforcement—that is, the willingness of people to incur a cost to punish a free rider[1]—can potentially explain the maintenance of high levels of cooperation (Güth et al. 1982; Fehr and Gächter 2000, 2002). For instance, a decentralized norm enforcement mechanism can successfully prevent the exploitation of common pool resources (Ostrom 1990), whereas within firms, social pressure and mutual monitoring can play an important role in inducing effort and thereby contribute crucially to an organization's efficiency (Kandel and Lazear 1992, Mas and Moretti 2009). However, if punishment is used indiscriminately, its cooperation-enhancing effect may vanish, and large welfare losses may result (Herrmann et al. 2008).

This paper investigates in a laboratory experiment how private monetary incentives for contributing to a public good affect decentralized norm enforcement through peer punishment. In our experiment, both the incentives and the peer punishment have consequences for the monetary payoffs of the subjects, but whereas the private incentives are provided by a central authority (the experimenters in this case), the peer punishment is carried out by the individual group members after observing everybody's contribution to the public good, and is costly to both the punisher and the receiver.

Monetary incentives are widely used by policy makers and managers to foster prosocial behavior. For example, recycling or the use of environmentally friendly technologies (such as hybrid cars) is

[1] Throughout the paper, we use the term "free rider" to refer to individuals whose contributions to the public good are below average.

often subsidized. Within organizations, teamwork, contributions to the work atmosphere, and extra-role behavior are often encouraged with private monetary incentives such as bonuses, awards, or promotions. Offering private incentives to behave prosocially reduces the cost of prosocial behavior for contributors, and this has been found to be effective in raising contributions in a number of situations, as standard economics would predict.[2] However, a growing body of evidence in psychology (for a survey, see Deci et al. 1999) and economics (for surveys, see Frey and Jegen 2001, Bowles 2008) finds that monetary incentives can crowd out individuals' willingness to behave prosocially, leading to a *direct* detrimental effect of monetary incentives. Our paper focuses instead on the question of whether private monetary incentives, such as bonuses for extra-role behavior in a firm, can have an *indirect* effect on the level of prosocial behavior by affecting the functioning of a norm enforcement mechanism.

How private monetary incentives affect overall prosocial behavior in a setting where norm enforcement is important depends on how the incentives affect two crucial factors of norm enforcement: the propensity of prosocial individuals to inflict costly punishment on free riders, and the reaction of free riders to such punishment. If the incentives do not influence those factors, the effects of the incentives and of norm enforcement will be additive and a combination of the two should be most successful in fostering high levels of prosocial behavior. However, as we will argue below, monetary incentives can dampen the effectiveness of the norm enforcement mechanism, leading to less punishment and, as a consequence, lower contribution rates—even with an additional incentive in place.

How could monetary incentives *negatively* influence the two factors of a successful norm enforcement mechanism? First, norm enforcement mechanisms depend on the willingness of some individuals to punish free riders. It seems that high contributors are motivated to punish free riders because doing so allows them to vent their anger and express disapproval (Bosman and van Winden 2002, Fehr and Gächter 2002, Xiao and Houser 2005, Ben-Shakar et al. 2007, Hopfensitz and Reuben 2009), and that they derive satisfaction from the act of punishing norm violations (de Quervain et al. 2004). An extra incentive for a prosocial individual can potentially mitigate his anger, because the advantage of free riding is reduced by the incentive reward. For example, in the absence of private incentives to exert effort that benefits the firm (and therefore, directly or indirectly, all its employees), hard-working employees may feel angry when observing free riders who receive the same salary, leading to norm enforcement efforts. However, if employees are rewarded for hard work and extra-role behavior, their willingness to punish may be reduced, because they receive something that free riders do not.

Second, a successful punishment mechanism also relies on free riders adapting their behavior in response to punishment. For instance, employees who do not contribute their "fair share" to the public goods in a firm may feel compelled to increase their contributions after being sanctioned by their peers. However, in the presence of private incentives for contributors, free riders may not feel guilty for not contributing, and perceive punishment as unjustified, because they already forgo the additional incentive. As a result, free riders may not increase their contributions after punishment as much as when no private incentives are in place. Thus, if private monetary incentives dampen one or both of the factors that together sustain prosocial behavior over time, the norm enforcement mechanism may not be as effective as when no private incentives are present, and this may lead to lower overall prosocial behavior.

In our experiment, participants play two six-period public good games in groups of four: one without and one with a punishment opportunity. The baseline treatment, which is closely based on earlier experiments in the literature, works as follows: In every period, each group member decides how much of his endowment to contribute to a "group project." Payoffs are such that it is in every individual's best interest not to contribute, even though the group as a whole is best off if everybody contributes their full endowment. In the game with the punishment opportunity, participants can in each period assign "deduction points" to the other group members after seeing their contributions; these points have a monetary cost to the sender, but an even larger cost to the receiver. In the treatment that is novel to this paper, we add a private monetary contribution incentive to this setting: For each unit a participant contributes to the group project, he receives a "lottery ticket" with a relatively substantial expected value (though it is still in his private interest not to contribute).

The results show that the presence of this private incentive can indeed negatively affect both factors of the norm enforcement mechanism. First, offering salient monetary incentives proportional to contributions leads to less severe punishment of free riders by the other members of their group. In the setting with monetary incentives, deviations from the group average are punished less harshly than in a setting

[2] For example, tax deductions have been shown to increase charitable giving (Auten et al. 2002), and in experimental studies, peoples' giving behavior reacts consistently to changes in prices (Andreoni and Miller 2002, Fisman et al. 2007).

where no monetary incentives are offered. Second, punishment has less influence on free riders' behavior when monetary incentives for contributors are in place. For each punishment point received, free riders increase their subsequent contribution by less when monetary incentives are offered than when no incentives are offered. As a result of the negative reactions of both factors to incentives, prosocial behavior is not increased by the nontrivial incentive we provide—even though the incentive increases contributions substantially in the absence of norm enforcement.

These findings indicate that policy makers and managers should be careful in using private incentives to foster prosocial behavior in settings where norm enforcement and social pressure are important. We therefore contribute to the discussion about possible "hidden costs" of incentives (Lepper and Greene 1978). A number of empirical studies find negative effects of incentives on individuals' prosocial behavior (for example, Frey and Oberholzer-Gee 1997; Gneezy and Rustichini 2000a, b; Mellström and Johannesson 2008), and a variety of potential channels have been suggested to explain such detrimental effects of incentives. Extrinsic incentives might destroy intrinsic motivations to behave prosocially (Deci 1975, Frey 1997), reduce trust in a principal-agent relationship (Fehr and Falk 2002, Fehr and List 2004, Falk and Kosfeld 2006), or shift an individual's decision frame from a social to a monetary frame, suggesting that selfish behavior is acceptable or even appropriate (Gneezy and Rustichini 2000a, Heyman and Ariely 2004). Recent theories and experimental evidence additionally suggest that monetary incentives negatively affect individuals' image motivation in situations where their contribution to the public good is visible to others (Benabou and Tirole 2006, Ellingsen and Johannesson 2008, Ariely et al. 2009).

All these studies suggest potential direct channels through which incentives can have detrimental effects on prosocial behavior. The contribution of our experiment, on the other hand, is to demonstrate the possible importance of an indirect channel: Private incentives may lead to less effective norm enforcement, which in turn may eliminate the positive effect of incentives on prosocial behavior. Meanwhile, the incentives do have a positive effect on contributions in the part of our experiment where the norm enforcement mechanism is not available. This means that the direct channels through which incentives may reduce prosocial behavior do not appear to be present in our laboratory setting, or that the incentives are large enough for a possible negative direct effect to get swamped by the positive "price effect" predicted by standard economics. Indeed, most previous studies find that only *small* incentives can have a negative net effect on prosocial behavior (Gneezy 2003). The indirect channel identified in the punishment condition of our experiment, however, can lead to a negative net effect even for private incentives of relatively substantial size, which further enhances the potential practical importance of our findings.

In summary, the results of our experiment show that centrally provided monetary incentives, which may have a strong positive effect in settings without decentralized norm enforcement, can have a weak or even negative effect on prosocial behavior when norm enforcement is important and, by itself, powerful. Having said that, it is not clear in our setting that the presence of incentives does not increase welfare, even though contributions are not increased—after all, if a similar level of prosocial behavior can be achieved with less (socially wasteful) norm enforcement, this is welfare enhancing. In our experiment, overall welfare is unaffected by the incentives, because the reduction in punishment they produce is offset by the cost of providing them. More generally, although norm enforcement is costly in the short run, in the long run it is mainly the threat of punishment that maintains high levels of contributions, whereas providing incentives continues to be costly. Therefore, whether the introduction of private incentives for prosocial behavior is desirable for a policy maker or manager will depend on the weights she assigns to contributions versus norm enforcement costs, on the cost of providing incentives, and on the horizon over which the policy is considered.

The remainder of this paper is structured as follows: In §2, we describe our experimental design and its two treatments, the baseline treatment and the incentive treatment. Section 3 discusses our behavioral hypotheses, namely, the various ways in which the presence of incentives could influence the different aspects of the norm enforcement mechanism. Section 4 then presents the results from our experiment, and §5 discusses these further. Finally, in §6, we briefly conclude and suggest topics for further research.

2. Experimental Design

Our experiment consists of a linear public good game (also known as "voluntary contribution mechanism") with two treatments comprising two parts each (see Table 1). In the *baseline treatment* (BT), subjects participate in two six-period public good games with and without punishment opportunity. In the *incentive treatment* (IT), a private monetary incentive is added to the BT.

2.1. The Baseline Treatment (BT)

Our baseline treatment closely follows Fehr and Gächter (2002), except that we use a "partner" design

Table 1 Treatments

	Baseline treatment (without private incentives)	Incentive treatment (with private incentives)
Without punishment (six periods)	15 groups of size 4	19 groups of size 4
With punishment (six periods)	15 groups of size 4	19 groups of size 4

(meaning that groups remain fixed within each game) instead of their "stranger" design (meaning that groups are reshuffled after each period). We chose to do so because repeated interaction with the same group members is arguably more realistic for most real-world applications. Participants first play six periods of a public good game in fixed groups of four. In each period, each group member $i \in \{1, 2, 3, 4\}$ receives an endowment of 20 experimental currency units (ECU) and can contribute an integer g_i $(0 \le g_i \le 20)$ to a public good (referred to as a group project). All group members decide simultaneously on their g_i in a period. The monetary payoff of each individual i from the group project in a period is given by

$$\pi_i = 20 - g_i + a \sum_{j=1}^{4} g_j, \tag{1}$$

where a is the marginal per-capita return (MPCR) from a contribution to the public good. In this experiment, as in Fehr and Gächter (2000, 2002) and many subsequent papers in this literature, a is set to equal 0.4. Hence, the private cost to an individual of contributing 1 ECU to the public good is 0.6 ECU, whereas the total benefit to his fellow group members is 1.2 ECU. This means that not contributing at all ($g_i = 0$) is the dominant action for each group member i in the stage game, whereas the total group payoff ($\sum_{i=1}^{4} \pi_i$) is maximized if all group members contribute their full endowment ($g_i = 20$).

After six periods without punishment, participants are rematched into new groups and play another six-period public good game with the same parameter values as in the first six periods, but with a peer punishment mechanism.[3] In each period, participants now receive an additional endowment of 10 ECU.[4] After participants make their contribution decision g_i, they are informed about the contribution of each other group member $j \neq i$, and are allowed to assign punishment points, p_{ij} $(0 \le p_{ij} \le 10)$, to the other group members.[5] The punishment points (neutrally labeled deduction points in the instructions) are costly to both the sender and the receiver. Each punishment point costs 1 ECU to the sender and 3 ECU to the receiver. However, the payoff-effective punishment costs imposed by the other group members on subject i, C_i, cannot exceed the first-stage payoff, π_i. C_i is therefore given by $C_i = \min(3 \sum_{j \neq i} p_{ji}, \pi_i)$. The overall payoff of subject i in a period of the public good game with punishment is then given by

$$\widehat{\pi}_i = 10 + \pi_i - C_i - \sum_{j \neq i} p_{ij}. \tag{2}$$

2.2. The Incentive Treatment (IT)

The incentive treatment is identical to the baseline treatment, except that the subjects are now given a private monetary incentive to contribute to the public good. To make the private incentive salient, participants receive a (virtual) "lottery ticket" for each ECU that they contribute to the public good. Each lottery ticket gives a 1% chance of winning an additional 20 ECU at the end of the experiment, so it has an expected value of 0.2 ECU.[6] For example, if a participant contributes 10 ECU to the public good, he or she will, in expectation, win 2 ECU. Thus, the expected private monetary payoff in the IT is increased by $0.2 g_i$ compared with the payoffs in the BT. This incentive is in place for both parts of the treatment, with and without punishment.

The monetary incentive is nontrivial, because, in expectation, it is equivalent to a reduction from 0.6 ECU to 0.4 ECU in the private cost of contributing an ECU to the public good. It is important to note, however, that this incentive does not alter the benefits that the other members of i's group derive from i's contribution, and that from a purely monetary perspective, it is still a dominant strategy in the stage game for each subject to not contribute anything to the public good, unless the subject is extremely risk loving.[7]

[3] Participants were only informed about the second six-period game once the first game was over. Because Fehr and Gächter (2000, 2002) and Herrmann et al. (2008) find that the order of the nonpunishment and the punishment part has no major effects on behavior, we opted to use only the more natural order.

[4] This was done to reduce the likelihood that a subject would refrain from punishment to avoid the risk of a negative payoff in a period. Many experiments on punishment behavior (including Fehr and Gächter 2002) give participants an additional endowment in the punishment condition.

[5] As is common in this literature, the group members are listed in random order after each contribution decision to avoid individual reputation effects.

[6] As discussed below, one period of each six-period game is randomly chosen for payment at the end of the experiment, so that the maximum number of lottery tickets a subject can obtain is 40 (2×20). The lotteries were conducted so that each subject's tickets were "additive," not independent, within each six-period block of the treatment. This means that having x tickets gives a one-time x% chance to win 20 ECU, rather than x independent 1% chances of winning.

[7] We believe it is very unlikely that a subject would be willing to give up $0.6x$ ECU for an x% chance of winning 20 ECU, which is what would be needed to make contributing (at least) x the dominant action.

2.3. Procedures

The experiments were conducted at the Computer Lab for Experimental Research at Harvard University, using the software z-Tree (Fischbacher 2007). Sixty subjects, the vast majority of them undergraduate students, participated in the BT and 76 in the IT. Participants received detailed written instructions with a number of control questions, and the experiment started only after all participants answered all the questions correctly. (The instructions distributed to the subjects can be found in the online appendix, available in the e-companion.[8]) At the end of a session, one period of each six-period game was randomly chosen to determine the final payoff (consisting of the monetary payoff, and in the IT, the number of lottery tickets), and (in the IT) the two lotteries were played out. Then, participants were paid their total earnings from the games, converted at a rate of 1 ECU = US$0.25, and a show-up fee of US$10.00. Average earnings for the experiment, which lasted approximately 80 minutes, were US$24.40.

Whether using one randomly chosen period for the final payoff affects behavior as compared with paying subjects for each period is an open methodological question, although it has been found not to matter in at least some experimental games (Laury 2005). In our experiment, paying only one period per game allows us to increase the stake for each decision, and in particular, to offer a nontrivial prize in the lottery in the IT (20 ECU = US$5). In turn, having higher stakes may change behavior as compared with a situation with lower stakes. Most other papers in the literature pay subjects for every period, but at much lower conversion rates (for instance, Herrmann et al. 2008 convert at the rate 1 ECU = US$0.03). This may explain why we find somewhat lower contribution and punishment levels than most other papers in this literature.

3. Behavioral Hypotheses

The main goal of this paper is to investigate how the presence of centrally provided private incentives to contribute to the public good affects the functioning of a decentralized norm enforcement mechanism (peer punishment) and the resulting level of public good contributions. Previous research indicates that the ability of a peer punishment mechanism to sustain or increase public good contributions depends largely on two factors: (1) how harshly subjects who contribute less than average (free riders) are punished, and (2) how those free riders adapt their contributions afterward.[9] If private incentives do not interact with either of these factors, we should expect that the IT yields higher contributions than the BT, with and without the punishment mechanism. This is because the private cost of contributing is reduced, which should bring contributions to a higher level (a standard price effect).[10] However, as we discuss below, the presence of private incentives may influence punishment behavior and the reaction of free riders, and as a result, these interactions could yield higher or lower overall contribution levels than would be expected from the price effect alone.

3.1. Punishment of Free Riders

A growing body of research has investigated the driving forces behind punishment behavior. One of the main findings is that negative emotions toward free riders are an important motivation, or perhaps even the main motivation, behind punishment. High contributors punish free riders to vent their anger and express their disapproval (Bosman and van Winden 2002, Fehr and Gächter 2002, Xiao and Houser 2005, Ben-Shakar et al. 2007, Hopfensitz and Reuben 2009), and seem to derive satisfaction from doing so (de Quervain et al. 2004).[11]

Depending on exactly what determines the strength of negative emotions toward free riders, monetary

[8] An electronic companion to this paper is available as part of the online version that can be found at http://mansci.journal.informs.org/.

[9] The extent of "antisocial punishment" (meaning the punishment of above-average contributors) is also crucial, as shown by Herrmann et al. (2008). However, because we do not expect (or find) much antisocial punishment in our American subject pool, we do not focus on it in our discussion.

[10] Of course, this already assumes that subjects do not act in accordance with standard game theory, which predicts that nobody ever contributes (or punishes). Various papers have looked at how contributions are affected by changes in the MPCR, a, in public good experiments without punishment, and they generally find a fairly strong and significant price effect (see the survey in Ledyard 1995). A recent paper by Carpenter et al. (2009), which varies a as well as group size in a stranger setting with punishment, finds that an MPCR of 0.75 rather than 0.3 leads to higher mean contributions when the group size is eight, but not when it is four. As discussed earlier, we do not change a across treatments, but rather give a personal incentive akin to a rebate in the IT.

[11] This explains why high contributors punish free riders even in a pure stranger design (where subjects are certain never to be in a group with the same other subjects again) (Fehr and Gächter 2002, Egas and Riedl 2008). If a group interacts repeatedly, strategic reasoning (namely, punishing free riders to lead them to increase their future contributions) provides another motivation for punishment. However, it seems that its importance is rather small as compared with the nonstrategic motivations. For instance, punishment patterns in a repeated public good game are very similar when punishment choices are revealed only at the end of the experimental session as opposed to after every period (Vyrastekova et al. 2008), and punishers do not seem to view another group member's punishing of a free rider as a substitute for their own punishing (Casari and Luini 2008).

incentives for contributing to the public good may either increase, have no effect on, or decrease punishment. To illustrate the three different possible effects of incentives on punishment, it may be helpful to consider the following simple one-shot, two-person social dilemma. First-stage payoffs are given by $\pi_i = E - (1-r)g_i + a \cdot (g_i + g_j)$, where E is the subjects' endowment, g_i the contribution of subject i, r the private reward a subject receives per unit of contribution, and a the MPCR of the public good.[12] Assume that $g_i > g_j$ and that we are interested in i's punishment decision. What determines i's anger toward j?

Case 1. The first possibility is that high contributors become angry at people they see as selfish, and want to punish them for their selfish traits, à la Levine (1998). In terms of our social dilemma, we could define j's "selfishness" s_j as the "benefit withheld from the other player" divided by the unit cost of providing this benefit, or

$$s_j = \frac{a(E - g_j)}{1 - r}, \tag{3}$$

and we can then imagine that i's punishment of j increases in j's selfishness as compared to i's,

$$s_j - s_i = \frac{a(g_i - g_j)}{1 - r}. \tag{4}$$

Clearly, this expression increases in r. Intuitively, not contributing (or contributing little) is more selfish the lower the personal cost of contributing. Thus, this would predict that there should be more punishment in the IT because not contributing is a more selfish action there.

Case 2. Alternatively, high contributors may punish based on the harm that the low contributors' actions impose on them, without taking into account the cost of contributing. In terms of our example, this would mean that i punishes j solely based on the numerator of the previous expression, $a(g_i - g_j)$, which is independent of r. For our experiment, where the MPCR a is the same in both treatments, this would predict equal strength of punishment in both treatments. This case would be consistent with models of intention-based reciprocity such as Rabin (1993), in which j's kindness toward i depends only on the consequences of j's actions for i's payoff.

Case 3. Finally, it may be that inequality of outcomes is what triggers negative emotions toward low contributors and therefore punishment. In our example, we have

$$\pi_j - \pi_i = (1 - r)(g_i - g_j), \tag{5}$$

which is decreasing in r, meaning that the payoff inequality for a given contribution difference is smaller if contributing is less costly. In our experiment, this would predict that punishment of free riders is less harsh in the IT than in the BT, because high contributors receive a private reward in the former but not in the latter.

Such behavior would be consistent with outcome-based theories of social preferences, such as the inequity-aversion theory of Fehr and Schmidt (1999). However, previous experiments have revealed that reducing payoff inequalities does not seem to be the main motive behind punishment, because punishment levels are substantial even when the cost is the same to the punisher as to the receiver, and therefore inequality is not reduced by punishment (Falk et al. 2005, Egas and Riedl 2008, Masclet and Villeval 2008). It is important to note, though, that even if the reduction of payoff inequality is not the main *goal* of punishment, it may still be the case that inequality is the *cause* of the negative emotions that lead to punishment.[13] This would be consistent with the findings of Dawes et al. (2007), who look at punishment in a setting where first-stage payoffs are randomly generated (in fact, drawn from the distribution of first-stage payoffs in Fehr and Gächter 2002) rather than determined by contribution decisions. Dawes et al. find that high earners are still punished substantially, even though their high earnings are not a result of free riding, and that in a hypothetical scenario, their subjects express negative emotions (annoyance and anger) toward high earners, the more strongly the higher the inequality.

In summary, under different assumptions of what drives punishment, we would predict different effects of private contribution incentives on the strength of punishment. Our results will therefore not only show whether private incentives interact with punishment behavior, but also give an indication of which assumption regarding the causes of negative emotions toward free riders is most reasonable (at least in the public good context used in our experiment).

3.2. Reaction of Free Riders

The success of a peer punishment mechanism in increasing contributions depends not only on sufficiently harsh punishment of free riders, but also on

[12] a and r are such that $(1-r)/2 < a < 1 - r$, which makes this a social dilemma, as it is in each subject's private monetary interest to choose $g_i = 0$, while total payoff $\pi_i + \pi_j$ is maximized if $g_i = g_j = E$.

[13] This is noted, for instance, by Fehr and Gächter (2004, p. E1) in their response to Fowler et al. (2004), who point out that "egalitarian motives," as opposed to negative emotions toward free riders, may be responsible for the punishment observed in Fehr and Gächter (2002): "Fowler et al. contrast their egalitarianism hypothesis with our view that negative emotions against free riders drive punishment. However, the two views are not necessarily incompatible: egalitarian sentiments may be the basis behind cooperators' negative emotions because free riding causes considerable inequalities."

their reaction to the punishment. It has been observed in numerous experiments that free riders who are punished in a period increase their contribution in the subsequent period. One reason for doing so is surely to avoid the material costs from further punishment. However, this may not be the only motivation: A free rider may also increase his contributions because he feels bad for not adhering to the contribution norm of his group, and this feeling may be enhanced if he is punished by the other group members, who thereby clearly signal their disapproval of the free rider's action. Bowles and Gintis (2005) refer to these two respective feelings as guilt and shame, and argue that these emotions play a significant role in increasing the contributions of free riders.

Direct evidence for the possible importance of guilt and shame in a public good experiment is provided by Masclet et al. (2003), who investigate the effect of nonmonetary sanctions. In their experiment, subjects can express their disapproval about the actions of another group member by assigning "disapproval points" without any monetary costs to either the sender or the receiver of these points. The results show that free riders who receive more disapproval points increase their contributions by more in the next period, consistent with the "shame" hypothesis. This is the case even in a stranger setting, where subjects are rematched into new groups in every period. Another finding is that free riders who are furthest below the average contribution in the previous period increase their subsequent contributions the most for a given level of punishment, a finding that is consistent with the "guilt" hypothesis.

Hopfensitz and Reuben (2009) provide further evidence by looking at a one-shot, two-person trust game with punishment and possible counterpunishment. Hopfensitz and Reuben measure various emotions (such as anger, guilt, shame, surprise) of the players directly through questionnaires right after players observe their partner's action, but before they choose their own action. They find that among the second movers who are punished for defecting (meaning that they returned little of the entrusted money), those who reported feeling guilty were less likely to retaliate than those who did not feel guilty, and they returned more money when playing the game a second time (against a different first mover). Furthermore, the intensity of guilt expressed by second movers seems independent of whether they were punished. These results support the claim that "prosocial emotions" such as guilt and shame are crucial for the effectiveness of a punishment institution.

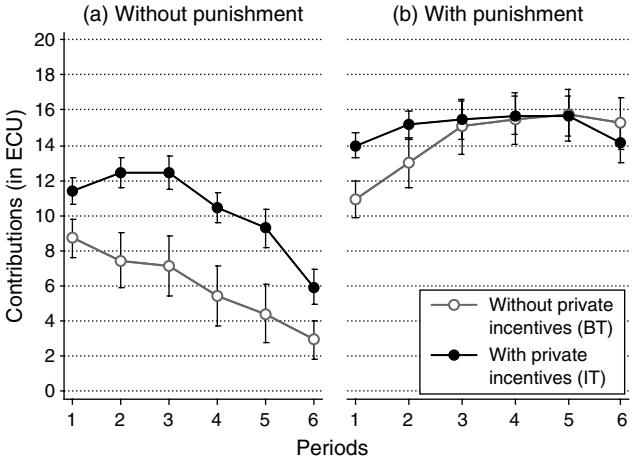
For our experiment, this means that the effect of private incentives on contributions may depend on how these incentives influence the extent to which free riders feel guilty or ashamed for not contributing as much as their peers. In ways similar to the ones discussed in the previous section, it is conceivable that the presence of private incentives enhances or reduces the strength of these emotions or leaves them unaffected. Thus, if private incentives increase the perceived selfishness of free riders, and as a consequence their guilt and shame as in Case 1, then the reaction of free riders to a given level of punishment will be stronger. However, if inequality drives not only the anger of the high contributors but also the guilt and shame of free riders (as in Case 3), then the presence of incentives will lower free riders' guilt and shame and will dampen their reaction to punishment.[14]

The overall effect of private incentives on the contribution level in the public good game will then depend on the combination of three factors: (1) the price effect of the incentive (it becomes cheaper to contribute); (2) the intensity of the punishment inflicted on free riders; and (3) the change in free riders' contribution behavior over time. If (2) and (3) are unaffected by the presence of incentives, we expect contribution levels to go up, in magnitude comparable to the incentive effect in a setting without peer punishment. If punishment becomes harsher or free riders react more strongly to it (in the sense of increasing their subsequent contributions by more), then the private incentives will have a larger positive effect on contribution levels than in a version of the game without punishment—in other words, incentives and peer punishment will be *complements*. However, if instead the presence of incentives leads to weaker punishment of free riders or if (punished) free riders are less prone to increase their contributions over time, the positive effect of incentives will be diminished, and if the negative effects of incentives on punishment and the free riders' reaction to it are sufficiently strong, contributions may even be lower than without incentives.[15] In that case, one could say that incentives and peer punishment are *substitutes*.

[14] A recent paper by Reuben and Riedl (2009) considers a different twist on the public good game with punishment and contains findings consistent with that last possibility. Reuben and Riedl experimentally investigate contributions and sanctioning in "privileged groups," meaning that for one group member it is privately optimal to contribute his full endowment to the public good, because the benefit he derives from it is sufficiently high. They find that such privileged groups obtain significantly higher contributions than "normal" groups when no peer punishment is possible, but that this is reversed with peer punishment. The main reason for this seems to be that the "low-benefit" subjects in privileged groups are less willing to increase their subsequent contributions in response to being punished than the low-benefit subjects in normal groups (where all group members are low-benefit).

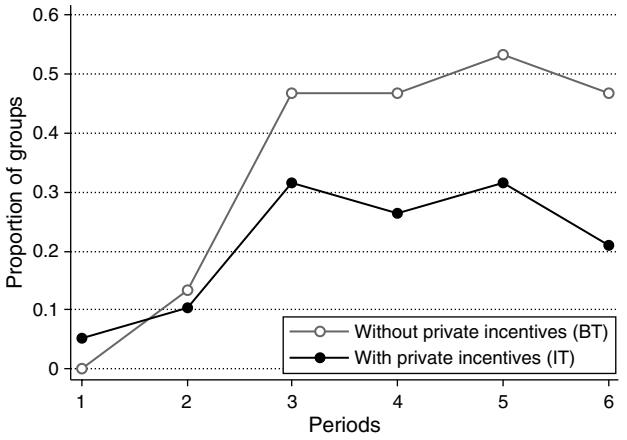
[15] Of course, it is possible that punishment becomes stronger while the reaction of free riders becomes weaker, or vice versa. In such a case, the total effect on contributions is ambiguous.

Figure 1 Effect of Incentives on Mean Contributions to Public Good



Note. Bars show standard errors of the group means.

Figure 2 Proportion of Groups Reaching Socially Efficient Contribution Level in the Punishment Condition



4. Results

4.1. Effect of Incentives on Contributions

Result 1. *Without the punishment mechanism, contributions are significantly higher in the IT than in the BT, whereas with the punishment mechanism, there is no significant difference between the two treatments.*

Support for Result 1 is presented in Figure 1. Figure 1(a) shows that in the six periods without the punishment mechanism, the presence of private incentives leads to significantly higher contributions. On average, contributions are 10.3 ECU with private incentives and only 6.0 ECU without incentives, and a nonparametric Mann-Whitney test rejects the null hypothesis of equal distributions of group average contributions ($z = -2.672$, $p < 0.01$).[16]

However, in the periods with the punishment mechanism, contributions are not significantly higher when a monetary incentive is added. Average contributions over all periods are 14.2 ECU in the BT and 15.0 ECU in the IT, and a Mann-Whitney test does not reject the null hypothesis of equal distributions $z = -0.173$, $p = 0.86$). As can be seen from Figure 1(b), contributions are higher in the IT for only the first two periods but are never statistically significantly so at the 95% level. If we look only at the last four periods, average contributions are actually slightly higher in the BT (15.4 ECU versus 15.2 ECU).

To gain a better understanding of the differences in contribution behavior between the two treatments when the punishment mechanism is in place, it is useful to compare the proportion of groups that reach the socially efficient outcome of everybody contributing 20 ECU in a period. It may be the case that average contributions in the IT are not significantly higher than in the baseline because of a "ceiling effect;" that is, that participants are unable to increase contributions beyond 20 even though they might be willing to do so if they could. In fact, however, the proportion of groups who reach the socially efficient outcome is *lower* in the IT than in the BT from period 2 onward (see Figure 2).[17] Thus, the result that private incentives do not increase contributions when a punishment mechanism is in place is not because of a ceiling effect.

To sum up, these results show that incentives have a significant positive effect on contributions when no peer punishment is possible. Thus, our "lottery ticket" incentives "work." However, with peer punishment, this is no longer the case; contribution levels are indistinguishable across the two treatments, and a somewhat higher proportion of groups manage to reach the efficient outcome in the BT than in the IT. The next two subsections compare punishment and subjects' reaction to it across the two treatments.

4.2. Effect of Incentives on Peer Punishment

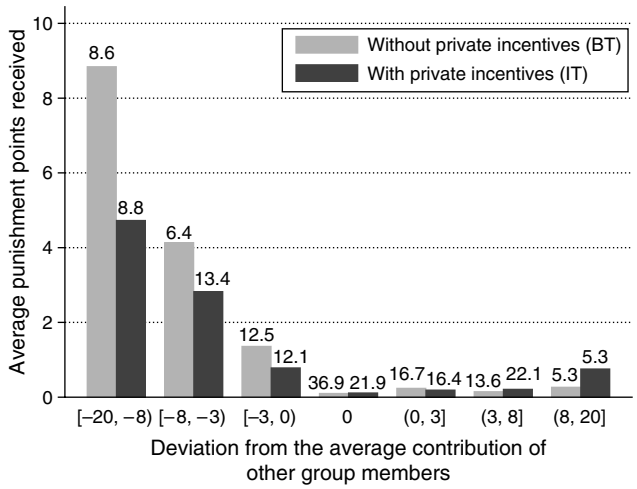
Result 2. *Group members who contribute less than average (free riders) are punished significantly less severely in the IT than in the BT.*

Figure 3 shows the average number of punishment points a subject received as a function of the deviation of the subject's contribution from the average contribution of the other three group members. In both treatments, free riders are punished, and the more strongly so the further their contributions are below the average. However, this increase in the severity of

[16] "Group average contribution" refers to the mean contribution level in a group across all six periods. Unless otherwise stated, we report two-sided test statistics.

[17] The proportions are not quite significantly different at conventional levels, though; the p-value from a one-sided Fisher's exact test is 0.11 for period 6 and higher for the other periods.

Figure 3 Received Punishment Points for Deviations from Others' Average Contribution



Note. Numbers above bars indicate the relative frequency of observations in a category.

punishment is much less pronounced in the IT than in the BT. On average, a free rider receives 4.34 punishment points in the BT, but only 2.59 points in the IT, and the difference is significant in a Mann-Whitney test ($z = 2.935$, $p < 0.01$). It is the severity of punishment of free riders that is different across treatments, not the frequency: In the BT, free riders are punished in 75.8% of cases, whereas in the IT they are punished in 71.2% of cases, and the difference is not statistically significant ($p = 0.26$, one-sided Fisher's exact test).

Meanwhile, the number of punishment points received by subjects who contribute more than average does not differ significantly across the two treatments (Mann-Whitney: $z = 0.241$, $p = 0.8$).[18]

Given the possible statistical dependence of observations due to the repeated appearance of the same subjects and groups in our data, the assumptions underlying the Mann-Whitney tests reported above may not be satisfied. We therefore resort to more sophisticated statistical techniques, which allow for such dependence, and furthermore enable us to disentangle what drives the differences between treatments. The Tobit regressions in Table 2, which cluster standard errors at the group level, confirm the impression from the graph and the earlier tests. We regress the number of punishment points a subject i receives in a period on the average contribution of the other group members and i's deviation from this average, allowing for different coefficients for positive and negative deviations, and also controlling for period effects. Both with and without incentives, the coefficient on negative deviations from the others' average is negative and highly significant, meaning that the farther a subject's contribution is below the average, the more the subject is punished. However, as can be seen in the last column, this effect is significantly stronger without incentives, confirming the impression from the graph. In terms of magnitudes, the predicted marginal effects are that an additional one-point negative deviation from the average leads to a 0.315 point increase in punishment in the BT, but only a 0.217 point increase in the IT.[19] Thus, the marginal effect of negative deviations on punishment is about 31% lower in the IT than in the BT.

Positive deviations from the average, on the other hand, do not significantly affect received punishment in either treatment. A higher average contribution of other group members is predicted to significantly

Table 2 Determinants of Punishment Points Received

	(1) Without incentives	(2) With incentives	(3) Pooled
Incentive			−3.038**
			(1.514)
Neg. deviation from others' avg.	−1.053***	−0.718***	−1.032***
	(0.076)	(0.112)	(0.072)
Incentive × neg. dev.			0.32***
			(0.12)
Pos. deviation from others' avg.	−0.137	−0.014	−0.174
	(0.089)	(0.119)	(0.111)
Incentive × pos. dev.			0.165
			(0.165)
Others' average contribution	−0.201***	0.037	−0.197**
	(0.078)	(0.063)	(0.087)
Incentive × others' avg.			0.233**
			(0.108)
Constant	−0.45	−2.462*	0.21
	(1.552)	(1.26)	(1.341)
Period dummies	Yes	Yes	Yes
No. of observations	360	456	816

Notes. Dependent variable: punishment points received by a subject. Tobit regressions. *Incentive* is a dummy variable that equals one for observations from the incentive treatment. *Neg. deviation from others' avg.* = $\min(0, g_i - \bar{g}_{-i})$; *Pos. deviation from others' avg.* = $\max(0, g_i - \bar{g}_{-i})$. Standard errors in parentheses clustered at the group level.

*$p < 0.1$; **$p < 0.05$; ***$p < 0.01$.

[18] The frequency of punishment of such subjects, which is 15.6% in the BT and 14.0% in the IT, is also statistically indistinguishable across the two treatments ($p = 0.26$, one-sided Fisher's exact test).

[19] The marginal effects refer to changes in the unconditional expected number of punishment points received, and are calculated from the "pooled" regression in the final column of the table, at the sample means for all values. The marginal effects predicted from columns (1) and (2) are very similar. The regressions in Table 2 are also robust to the inclusion of group dummies. In particular, the main coefficient of interest, *incentive × negative deviation*, is almost unchanged (0.28) and significant at $p < 0.05$. The predicted marginal effects on punishment from deviating an additional point from the group average are also very similar, at 0.284 and 0.199 (and thus 30% lower in the IT).

reduce punishment in the BT, but not in the IT (however, the predicted marginal effect is small in the BT—a one-point increase in others' average contribution is predicted to reduce punishment received by 0.06 points). Finally, note that in the pooled regression, the incentive dummy is negative and significant, meaning that there is less punishment in the IT, controlling for the others' average and deviations from the average.[20]

Another way to appreciate the quantitative difference between the two treatments in the intensity of punishment of free riders is to look at the predicted punishment received by a hypothetical subject. Assume that a subject contributes 10 ECU, whereas the three other group members contribute, on average, 15 ECU. The coefficients from the Tobit regression then predict that the subject receives 2.28 punishment points in the BT, but only 1.57 punishment points in the IT. If the hypothetical subject contributes nothing, he is predicted to receive 11.92 punishment points in the BT and only 7.59 in the IT.

4.3. Reaction of Free Riders to Punishment

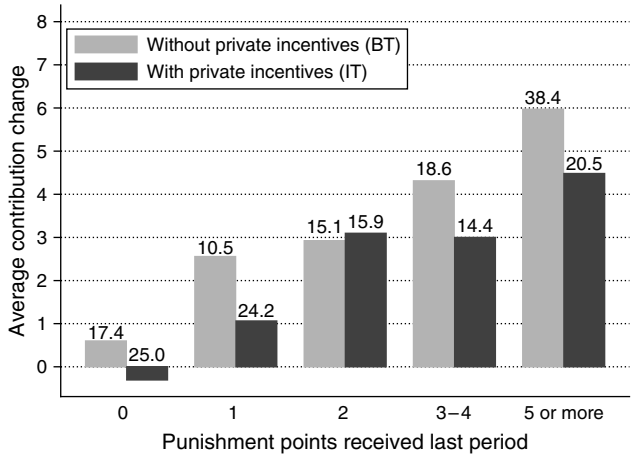
Result 3. *For a given level of punishment, free riders increase their subsequent contributions by less in the IT than in the BT. This seems mostly due to the unwillingness of free riders in the IT to increase their contribution toward the average of their fellow group members.*

Again, we provide both graphical and regression support for the result. Figure 4 displays the changes in contributions from one period to the next for free riders (subjects who contributed less in the previous period than the other group members did on average) who received different numbers of punishment points. Clearly, free riders increase their contributions by more if they are punished more heavily. However, the figure also shows that free riders tend to increase their contributions by less in the IT than in the BT. On average, free riders in the BT increase their contribution in the next period by 3.91 points, whereas in the IT, the increase is only 2.02 points, and this difference is statistically significant at $p < 0.01$ (Mann-Whitney test, $z = 2.767$). Looking only at those free riders who were punished, the average increases were 4.61 and 2.80 points ($z = 2.190$, $p < 0.03$).

To disentangle what drives the differences in the behavior of free riders between the two treatments, we look at regressions of the change in contribution on the severity of punishment and other explanatory variables such as the average contribution of other group members in the previous period and the free rider's deviation from it. This again allows us to account for possible statistical dependence of observations within groups and due to the repeated appearance of the same subjects.

[20] However, this coefficient may simply compensate for the absence of a negative coefficient on others' average in the IT.

Figure 4 Contribution Change of Free Riders in Punishment Condition



Note. Numbers above bars indicate the relative frequency of observations in a category.

The first column of Table 3 shows the coefficients of a regression of the change in contribution, $g_{i,t} - g_{i,t-1}$, on the cost of received punishment in the previous period, $C_{i,t-1}$, and the deviation of the subject's contribution in the previous period from the average of the other subjects' contributions, $g_{i,t-1} - \bar{g}_{-i,t-1}$ (which are all negative, given that we include only free riders in the regression). We interact the explanatory

Table 3 Determinants of Free Riders' Contribution Changes

	(1)	(2)	(3)
Incentive	−1.489*	2.201	1.448
	(0.78)	(1.529)	(1.743)
Cost of punishm. received last pd.	0.156**	0.134***	0.17***
	(0.07)	(0.049)	(0.062)
Incentive × cost of punishment	−0.013	0.064	−0.024
	(0.099)	(0.07)	(0.098)
Neg. deviation from others' avg. last pd.	−0.051		0.124
	(0.22)		(0.25)
Incentive × neg. dev. last pd.	−0.132		−0.305
	(0.266)		(0.29)
Others' average contribution last pd.		0.263***	0.294***
		(0.071)	(0.101)
Incentive × others' avg. last pd.		−0.297***	−0.31**
		(0.109)	(0.135)
Constant	0.99	−1.463	−1.642
	(0.618)	(0.99)	(1.028)
Period dummies	Yes	Yes	Yes
No. of observations	218	218	218

Notes. Dependent variable: change in contribution. Linear (OLS) regressions. Standard errors in parentheses clustered at the group and the individual levels, following Cameron et al. (2006) and using their "cgmreg" routine in Stata.

*$p < 0.1$; **$p < 0.05$; ***$p < 0.01$.

variables with an IT dummy to detect differences between the two treatments. As in the previous section, we also control for period effects by including period dummies. Standard errors are clustered on groups and individuals.

The results of the first column show that free riders increase their contributions significantly more the more heavily they were punished in the previous period. The coefficient of 0.156 means that for each punishment point a free rider receives, he or she on average increases his subsequent contribution by about 0.47 points, because each punishment point received costs him 3 ECU (unless he is punished so heavily that his whole first-stage payoff is lost). The coefficient on the interaction of the incentive dummy and the cost of punishment is very small and insignificant, meaning that the marginal effect of punishment on subsequent contribution changes is equally strong in both treatments. However, the coefficient on the incentive dummy itself is negative and mildly significant ($p = 0.056$), meaning that for a given level of punishment and deviation from the average, free riders in the IT tend to increase their contributions by less than free riders in the BT do. On the other hand, in neither treatment does it seem to matter how much below the others' average a free rider's contribution was in the previous period. This is somewhat surprising, because, for instance, Masclet et al. (2003) find a large and significant coefficient on this variable when running a similar regression.

The second column uses the other group members' average contribution, $\bar{g}_{-i,t-1}$, as an explanatory variable, instead of i's deviation from this average. The coefficient on the cost of punishment is slightly lower, but even more highly significant, and there is still no significant difference across the two treatments. However, the coefficient on the incentive dummy has switched sign and is now positive. This is because the coefficient on others' average contribution is strongly significantly positive for the BT but very close to zero for the IT. Thus, the difference between the two treatments can be explained by the fact that in the BT, free riders increase their contribution more the more other group members contributed in the previous period, whereas the same is not true in the IT.[21]

Column (3) confirms these findings by reintroducing the free rider's lagged negative deviation from the average as an explanatory variable, which still does not enter the regression significantly, and leaves the other coefficients largely unchanged as compared with the results shown in column (2).[22]

4.4. Effect of Incentives on Welfare

RESULT 4. *Mean welfare is not significantly different across the punishment conditions of the two treatments, once the cost of providing the incentives is taken into account. Without the punishment mechanism, mean welfare is significantly higher in the IT than in the BT.*

Even though the private incentives provided in the IT fail to lead to higher contributions when peer punishment is possible, this does not mean that they are not welfare enhancing. This is because achieving a certain contribution level with less peer punishment is a good thing, as punishment is socially wasteful. Furthermore, the subjects in our IT are better off than the ones in the BT because we give them a lottery ticket for each point they contribute.[23] However, to assess the overall welfare effect of introducing incentives, one must take into account the cost of providing them, so the value of the lottery tickets should not enter the welfare calculation.[24] Therefore, we use expression (2) to compare welfare across treatments. Using this criterion, mean welfare is slightly, but not significantly, higher in the punishment condition of the IT than in the punishment condition of the BT (the respective mean payoffs, not including lottery tickets, are 35.0 ECU in the IT and 33.5 ECU in the BT; Mann-Whitney $z = 1.42$, $p > 0.15$). This is mostly due to the first two periods, during which contributions are higher in the IT than in the BT and the still-numerous free riders in the BT are punished harshly. Looking at

[21] To assess the difference between the two treatments in this regression, it is helpful to consider whether the predicted contribution change absent any punishment is higher in the BT than in the IT, as a function of others' average contribution. It turns out that the predicted contribution change is significantly higher (at $p \leq 0.05$) for $\bar{g}_{-i,t-1} \geq 11.2$, which is the case for more than 70% of free riders. At the median value of $\bar{g}_{-i,t-1}$, which is 14.5, the regression predicts that a free rider in the IT changes his contribution upward by 2.11 points less than a free rider in the BT.

[22] Because of the negative (but insignificant) coefficient on the interaction of the incentive dummy and $g_{i,t-1} - \bar{g}_{-i,t-1}$, the exercise conducted in the previous footnote now leads to predicting somewhat smaller and less significant differences between the two treatments. At median values for free riders of the two explanatory variables other than cost of punishment (which is again assumed to equal zero), the regression coefficients from column (3) predict that a free rider in the IT changes his contribution upward by 1.52 points less than a free rider in the BT, and the p-value of this difference is $p = 0.085$. If we drop the period dummies (which are jointly insignificant at a 5% significance level) and the interaction of incentive and cost from punishment, the predicted difference is 1.82 points and is significant at $p < 0.01$.

[23] The mean expected payoff (taking the value of a lottery ticket to be 0.2 ECU, its expected value) of a subject in the punishment condition of the IT was 38.0 ECU, as compared with 33.5 ECU in the BT (Mann-Whitney $z = 4.55$, $p < 0.001$).

[24] In the experiment, we (the experimenters) finance the incentive. However, in real-world situations, it would have to be financed through taxes (in case of incentives provided by the government) or directly by the party that is interested in raising contributions. This might lead to an additional welfare cost (because of deadweight losses from taxation, for instance).

periods 3–6 only, mean welfare is almost equal across the two treatments (36.4 ECU in the BT versus 36.0 ECU in the IT). As mentioned earlier, mean contributions are slightly higher in the BT during these periods; also, total punishment during these periods is almost the same in the two treatments. However, it is important to note that the result discussed in §4.2, namely, that free riders are punished more harshly in the BT than in the IT, still holds for these periods—there is no difference in total punishment, however, because there are fewer free riders in the BT.

For the no-punishment condition, mean payoffs exclusive of lottery tickets are 26.2 ECU in the IT and 23.6 ECU in the BT ($z = 3.78$, $p < 0.001$) (recall that in these rounds, subjects did not receive an additional 10 ECU lump sum in each period as in the punishment condition). Thus, unlike in the punishment condition, the introduction of lottery tickets does increase mean welfare when the punishment mechanism is not available.

5. Discussion

The main findings discussed in the previous section support the hypothesis that incentives and norm enforcement are substitutes, meaning that one or the other in isolation is successful in raising contributions, whereas adding incentives in a setting with a peer punishment mechanism does not lead to higher contributions. We find this to be due to two effects: (1) free riders receive significantly less punishment when incentives to contribute are present, and (2) they increase their subsequent contributions by less, whether or not they are punished.

Our preferred interpretation of the first effect is that the rewards (in the form of lottery tickets) received by the high contributors dampen their anger toward the free riders, resulting in less punishment. Thus, in terms of our discussion in §3.1, we find support for Case 3, namely, that lessening the inequality of outcomes by providing incentives reduces the anger of high contributors. The second effect is likewise consistent with the idea that the presence of contribution incentives mitigates the shame or guilt of free riders.

It may be debatable to what extent either (or both) of these explanations for the two effects is more compelling than explanations based on strategic reasoning. In particular, it is possible that the reduction in punishment is because of punishers' anticipating that free riders will not adjust their contributions upward very much. Similarly, free riders may react less strongly to punishment in the IT than in the BT, not because they feel less shame or guilt, but because they anticipate that high contributors will not punish them very harshly if they keep contributing less than average. Although our data do not allow us to rule out these "strategic" explanations for what we observe, we can look at what happens in and after the first period when the punishment mechanism is available to the agents. If less punishment was inflicted on free riders in the IT because these free riders react less strongly to punishment than when no incentives are available, we might expect that in the first period punishment severity would be similar in both conditions.[25] However, in our data, a free rider in the BT receives on average 6.2 punishment points in the first period, whereas in the IT the corresponding average is only 3.7 points (Mann-Whitney test: $z = 2.620$, $p < 0.01$). Unless high contributors in the IT somehow foresee that punishing free riders is "not worth it," which we think is unlikely, this observation means that "dampened anger" seems to be a better explanation than strategic considerations for the less harsh punishment of free riders in the IT. This interpretation is also in line with previous research, mentioned in §3.1, which finds that strategic explanations have rather low explanatory power for punishment behavior in such experiments.

As for the reaction of free riders, it is harder to dismiss the possibility that the weaker reaction of free riders (whether or not they are punished) is due to strategic reasoning, particularly because, as just mentioned, punishment is less harsh in the IT from the beginning. Thus, even though the contribution increases of free riders from one period to the next are lower in the IT from the beginning (they increase their contribution by 2.9 points on average between the first and the second period of the punishment condition, whereas the corresponding number in the BT is 5.7 points (Mann-Whitney $z = 1.91$, $p < 0.06$)), we do not know whether this is a result of reduced guilt/shame or whether they anticipated that failing to increase their contributions would not result in harsh punishment.

Another alternative explanation for what we observe is that the presence of incentives makes contributing seem less a "social act" and more an individual choice motivated at least partially by private benefits; this could also explain why free riders are not induced to increase their contributions toward the mean in the IT, whereas in the BT they are. A free rider in the IT may think that his fellow group members contribute a lot only because they are after the lottery tickets, not because they genuinely care about the well-being of the group, and therefore may feel no compunction about contributing less. In addition, such a subject may feel that punishment he receives from high contributors is unjustified, and he may

[25] Then, over time the (potential) punishers would realize that the free riders do not react much to punishment, and would reduce their punishment accordingly.

therefore refrain from increasing his subsequent contributions out of spite or principle. Again, we cannot rule out this explanation as an alternative to our story, which focuses more directly on the effect of incentives on emotions as motivators of behavior. However, we believe that this explanation fails to explain why high contributors punish free riders less harshly, unless one assumes that high contributors to some extent engage in "self-signaling" and infer their own motivation from their actions and the environment.[26] Furthermore, in the condition with no punishment mechanism, the lottery tickets perform well in increasing contributions, so there is no indication that they directly crowd out subjects' intrinsic motivation to contribute.

As discussed in §4.4, mean welfare is not significantly different across the two treatments when the punishment mechanism is available. The subjects gain on average from the presence of the incentives, because less socially wasteful punishment takes place, whereas average contributions are at the same level as without incentives. However, this positive effect on welfare is offset by the cost of providing the incentives. More generally, harsh punishment of free riders, which has a high social cost in the short run, can be expected to lead to high contributions in the long run mainly due to its effect as a threat, not because it is actually exercised. Thus, over time, we would expect a decline in the differences in welfare costs because of punishment across treatments (Gächter et al. 2008). Meanwhile, the private incentives must be provided in each period, and if their provision is socially costly (for instance, because of deadweight losses arising in their financing), then welfare may be higher without them. Of course, we cannot claim from our experimental results that this is what would actually happen in any given situation, but we believe that this possibility, which arises as a result of the detrimental effect of incentives on the effectiveness of the norm enforcement mechanism, should at least be considered by a policy maker or manager in deciding whether to introduce (additional) private incentives for prosocial behavior.

6. Conclusion

In our laboratory public good experiment, we find that private incentives for prosocial behavior, which substantially increase contributions in the condition without norm enforcement, fail to do so when norm enforcement is possible. This is because of the effects of the incentives on the severity with which free riders are punished, and on free riders' subsequent reaction. Our preferred interpretation of these findings is that being rewarded for their contributions reduces the anger of high contributors toward free riders, and that free riders may feel less shame or guilt for failing to contribute their fair share.

Thus, we have identified another mechanism through which incentives can have unintended side-effects, so-called "hidden costs." While the existing literature has identified several ways in which monetary incentives could directly crowd out prosocial behavior, our finding can instead be seen as an indirect hidden cost, because it operates through reduced effectiveness of the norm enforcement mechanism. As such, it should be of concern for policy makers and managers who contemplate introducing private monetary incentives in settings where norm enforcement can be expected to play a significant role in generating high contributions to a public good.

Several questions related to our hypotheses and findings await further research. For instance, it would be interesting to know more about exactly what determines the strength of the negative emotions toward free riders, which in turn trigger punishment. In our interpretation, the presence of incentives weakens these negative emotions. This interpretation is consistent with inequality of outcomes as a driving force behind the negative emotions. It would be desirable to elicit these emotions more directly, either through questionnaires (as done, for instance, by Hopfensitz and Reuben 2009) or through physiological or neuroscientific measurement. Also, we believe that the framing of the incentives may be important for their effect on norm enforcement. We chose to make the incentive very salient, but it may be that results would be quite different if we had instead implemented a direct rebate, such that the cost of contributing decreases without an explicit reward for contributing. Likewise, it would be interesting to see what would happen if instead of rewards for contributing, fines for not contributing were introduced.

7. Electronic Companion

An electronic companion to this paper is available as part of the online version that can be found at http://mansci.journal.informs.org/.

Acknowledgments

The authors are grateful to the editor, the associate editor, two anonymous referees, Luke Coffman, Armin Falk, Ernst Fehr, Simon Gächter, Lorenz Goette, Judd Kessler, Ernesto Reuben, and seminar audiences at Harvard University, the University of Zurich, and the Economic Science Association conference in Washington, DC, for helpful comments

[26] Then, for any given difference in contribution between two group members, the subject who contributed more might feel relatively less strongly in the IT than in the BT that he is being more prosocial than the other subject.

and discussions, and to Benjamin Levinger for help in conducting the experiments. The views expressed in this paper are solely those of the authors and not necessarily those of the Federal Reserve Bank of Boston or the Federal Reserve System.

References

Andreoni, J., J. Miller. 2002. Giving according to GARP: An experimental test of the consistency of preferences for altruism. *Econometrica* **70**(2) 737–753.

Ariely, D., A. Bracha, S. Meier. 2009. Doing good or doing well? Image motivation and monetary incentives in behaving prosocially. *Amer. Econom. Rev.* **99**(1) 544–555.

Auten, G., H. Sieg, C. T. Clotfelder. 2002. Charitable giving, income and taxes: An analysis of panel data. *Amer. Econom. Rev.* **92**(1) 371–382.

Benabou, R., J. Tirole. 2006. Incentives and prosocial behavior. *Amer. Econom. Rev.* **96**(5) 1652–1678.

Ben-Shakar, G., G. Bornstein, A. Hopfensitz, F. van Winden. 2007. Reciprocity and emotions in bargaining using physiological and self-report measures. *J. Econom. Psych.* **28**(3) 314–323.

Bosman, R., F. van Winden. 2002. Emotional hazard in a power-to-take experiment. *Econom. J.* **112**(476) 147–169.

Bowles, S. 2008. Policies designed for self-interested citizens may undermine "the moral sentiments": Evidence from economic experiments. *Science* **320**(5883) 1605–1609.

Bowles, S., H. Gintis. 2005. Prosocial emotions. L. Blume, S. Durlauf, eds. *The Economy as an Evolving Complex System, III*. Oxford University Press, New York.

Cameron, A. C., J. B. Gelbach, D. L. Miller. 2006. Robust inference with multi-way clustering. NBER Technical Working Paper 327, National Bureau of Economic Research, Cambridge, MA.

Carpenter, J., S. Bowles, H. Gintis, S.-H. Hwang. 2009. Strong reciprocity and team production: Theory and evidence. *J. Econom. Behav. Organ.* **71**(2) 221–232.

Casari, M., L. Luini. 2008. Peer punishment in teams: Expressive or instrumental choice? Working paper, Purdue University, West Lafayette, IN.

Dawes, C. T., J. H. Fowler, T. Johnson, R. McElreath, O. Smirnov. 2007. Egalitarian motives in humans. *Nature* **446**(7137) 794–796.

de Quervain D. J.-F., U. Fischbacher, V. Treyer, M. Schellhammer, U. Schnyder, A. Buck, E. Fehr. 2004. The neural basis of altruistic punishment. *Science* **305**(5688) 1254–1258.

Deci, E. L. 1975. *Intrinsic Motivation*. Plenum Press, New York.

Deci, E. L., R. Koestner, R. M. Ryan. 1999. A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. *Psych. Bull.* **125**(6) 627–668.

Egas, M., A. Riedl. 2008. The economics of altruistic punishment and the maintenance of cooperation. *Proc. Roy. Soc. B: Biol. Sci.* **275**(1637) 871–878.

Ellingsen, T., M. Johannesson. 2008. Pride and prejudice: The human side of incentive theory. *Amer. Econom. Rev.* **98**(3) 990–1008.

Falk, A., M. Kosfeld. 2006. The hidden cost of control. *Amer. Econom. Rev.* **96**(5) 1611–1630.

Falk, A., E. Fehr, U. Fischbacher. 2005. Driving forces behind informal sanctions. *Econometrica* **73**(6) 2017–2030.

Fehr, E., A. Falk. 2002. Psychological foundations of incentives. *Eur. Econom. Rev.* **46**(4–5) 287–324.

Fehr, E., S. Gächter. 2000. Cooperation and punishment in public goods experiments. *Amer. Econom. Rev.* **90**(4) 980–994.

Fehr, E., S. Gächter. 2002. Altruistic punishment in humans. *Nature* **415**(6868) 137–140.

Fehr, E., S. Gächter. 2004. Egalitarian motive and altruistic punishment (reply). *Nature* **433**(7021) E1–E2.

Fehr, E., J. A. List. 2004. The hidden costs and returns of incentives—Trust and trustworthiness among CEOs. *J. Eur. Econom. Assoc.* **2**(5) 743–771.

Fehr, E., K. M. Schmidt. 1999. A theory of fairness, competition, and cooperation. *Quart. J. Econom.* **114**(3) 817–868.

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

Fisman, R., S. Kariv, D. Markovits. 2007. Individual preferences for giving. *Amer. Econom. Rev.* **97**(5) 1858–1876.

Fowler, J. H., T. Johnson, O. Smirnov. 2004. Egalitarian motive and altruistic punishment. *Nature* **433**(7021) E1–E2.

Frey, B. S. 1997. *Not Just for the Money: An Economic Theory of Personal Motivation*. Edward Elgar, Cheltenham, UK.

Frey, B. S., R. Jegen. 2001. Motivation crowding theory: A survey of empirical evidence. *J. Econom. Surveys* **5**(5) 589–611.

Frey, B. S., F. Oberholzer-Gee. 1997. The cost of price incentives: An empirical analysis of motivation crowding-out. *Amer. Econom. Rev.* **87**(4) 746–755.

Gächter, S., E. Renner, M. Sefton. 2008. The long-run benefits of punishment. *Science* **322**(5907) 1510.

Gneezy, U. 2003. The *W* effect of incentives. Working paper, University of Chicago Graduate School of Business, Chicago.

Gneezy, U., A. Rustichini. 2000a. A fine is a price. *J. Legal Stud.* **29**(1) 1–18.

Gneezy, U., A. Rustichini. 2000b. Pay enough or don't pay at all. *Quart. J. Econom.* **115**(3) 791–810.

Güth, W., R. Schmittberger, B. Schwarze. 1982. An experimental analysis of ultimatum bargaining. *J. Econom. Behav. Organ.* **3**(4) 367–388.

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

Heyman, J., D. Ariely. 2004. Effort for payment: A tale of two markets. *Psych. Sci.* **15**(11) 787–793.

Hopfensitz, A., E. Reuben. 2009. The importance of emotions for the effectiveness of social punishment. *Econom. J.* **119**(540) 1534–1559.

Kandel, E., E. P. Lazear. 1992. Peer pressure and partnerships. *J. Political Econom.* **100**(4) 801–817.

Laury, S. K. 2005. Pay one or pay all: Random selection of one choice for payment. Working paper, Georgia State University, Atlanta.

Ledyard, J. O. 1995. Public goods: A survey of experimental research. J. H. Kagel, A. E. Roth, eds. *Handbook of Experimental Economics*. Princeton University Press, Princeton, NJ, 111–194.

Lepper, M. R., D. Greene, eds. 1978. *The Hidden Costs of Reward: New Perspectives on the Psychology of Human Motivation*. Erlbaum, Hillsdale, NY.

Levine, D. K. 1998. Modeling altruism and spitefulness in experiments. *Rev. Econom. Dynam.* **1**(3) 593–622.

Mas, A., E. Moretti. 2009. Peers at work. *Amer. Econom. Rev.* **99**(1) 112–145.

Masclet, D., M.-C. Villeval. 2008. Punishment and inequality: A public good experiment. *Soc. Choice Welfare* **31**(3) 475–502.

Masclet, D., C. Noussair, S. Tucker, M.-C. Villeval. 2003. Monetary and nonmonetary punishment in the voluntary contributions mechanism. *Amer. Econom. Rev.* **93**(1) 366–380.

Mellström, C., M. Johannesson. 2008. Crowding out in blood donation: Was Titmuss right? *J. Eur. Econom. Assoc.* **6**(4) 845–863.

Ostrom, E. 1990. *Governing the Commons: The Evolution of Institutions for Collective Action*. Cambridge University Press, Cambridge, UK.

Rabin, M. 1993. Incorporating fairness into game theory and economics. *Amer. Econom. Rev.* **83**(5) 1281–1302.

Reuben, E., A. Riedl. 2009. Public goods provision and sanctioning in priviledged groups. *J. Conflict Resolution* **53**(1) 72–93.

Vyrastekova, J., Y. Funaki, A. Takeuchi. 2008. Strategic vs. non-strategic motivations of sanctioning. CentER Discussion Paper 2008-48, Tilburg University, Tilburg, The Netherlands.

Xiao, E., D. Houser. 2005. Emotion expression in human punishment behavior. *Proc. Natl. Acad. Sci. USA* **102**(20) 7398–7401.