Economic Incentives and Social Preferences: A preference-based Lucas Critique of public policy

Samuel Bowles[§] and Sandra Polanía Reyes*

DRAFT 30th March, 2009

DO NOT CITE

Abstract

Laws and policies designed on the assumption that individuals have only self-regarding preferences may fail. This may occur when conventional self-interest-based policies diminish the salience of social preferences such as altruism and ethical predispositions to adhere to social norms and punish those who do not. Evidence from 50 experiments indicates that this so-called crowding out effect is pervasive, and that crowding in also occurs. A model of categorical and marginal crowding out is developed and evidence for the mechanisms underlying the non-additivity of self interested and social motives is provided. Incentives that appeal to self-interest may frame decision situations in a way that favors pursuit of self-interest or that conveys the information that a principal has a low opinion of an agent, or has selfish intentions, or believes that the agent's task is onerous. Explicit incentives may also diminish the signaling value of generous acts, or compromise an individual's sense of autonomy, thereby reducing the salience of intrinsic motives. Further, incentives alter social interactions in ways that influence the process by which individuals update their preferences and may lead to the abandonment of preexisting social preferences. Explicit incentives may also crowd in social preferences. All of these motivational crowding mechanisms reflect the fact that when individuals engage in production and exchange they have constitutive as well as acquisitive interests: they are interested not only in getting things, but also becoming and affirming that they are a particular kind of person. Public policy implementation thus cannot assume that preferences and beliefs are invariant to the taxes, subsidies or other instruments under consideration, but rather must evaluate policy consequences by considering outcomes that can be sustained once the endogenous nature of motivations is accounted for. The result is a preference-based analogue to the Lucas Critique of macroeconomic policy.

JEL codes: D64 (Altruism); H41 (Public goods); D78 (Policy making and implementation); Z13 (Social norms and social capital); C90 (Experiments)

Keywords: Public goods, behavioral experiments, social preferences, second best, motivational crowding

Affiliations: [§] corresponding author, Santa Fe Institute and University of Siena; * University of Siena. We thank Sung-Ha Hwang [acknowledgements to be completed] for their contributions to this paper and the Behavioral Sciences Program of the Santa Fe Institute, the U.S. National Science Foundation, the University of Siena and the European Science Foundation for support of this project.

1. Introduction

Thomas Schelling recalls his 'exciting and stimulating times' in the early 1950s as a young staffer at the White House in the Executive Office of the President. “People worked long hours,” he remembered in a recent communication to one of us, “and felt compensated by the sense of accomplishment, and ... personal importance. Regularly a Friday afternoon meeting would go on until 8 or 9, when the chairman would suggest resuming Saturday morning. Nobody demurred. We all knew it was important, and we were important. ... What happened when the President issued an order that anyone who worked on Saturday was to receive overtime pay and register in advance? Saturday meetings virtually disappeared.”

Since Richard Titmuss’ *The Gift Relationship: From Blood Donations to Social Policy,* economists have been intrigued but for the most part unpersuaded by the claim that policies based on explicit economic incentives may be counter-productive when they induce people to adopt a ‘market mentality’ and thus compromise pre-existing values to act in socially beneficial ways (Solow (1971), Arrow (1972b), Arrow (1972a), Bliss (1972)). At the time of its publication there were two strong reasons to doubt Titmuss’ claim. First, there was little hard evidence that the social preferences such as altruism, fairness, and civic duty that are said to be eclipsed by economic incentives are important influences on individual behavior. Second, even if these social preferences were thought to be important influences on behavior, there was even less evidence (in Titmuss’ book or elsewhere) that explicit economic incentives undermine them. As a result it was not implausible to hope that the social preferences and self-interest might contribute additively to the implementation of desirable social outcomes, or even in complementary ways. Thus one could agree with Arrow (1971) that “norms of social behavior, including ethical and moral codes (may) ...compensate for market failures” and not worry that explicit economic incentives designed to overcome market failures might compromise social norms and hence be ineffective or even counter-productive.

Theoretical and empirical advances over the intervening years provide the basis for a reconsideration of these issues. First, evidence from both the behavioral experimental laboratory and the field has demonstrated that social preferences are important influences on economic behavior (Bewley (1999), Fehr and Gachter (2000), Frey and Jegen (2001), Young and Burke (2001), Fehr and Falk (2002), Camerer and Fehr (2004), Fehr, Klein and Schmidt (2007)).

Second, the importance of incomplete contracts has been widely recognized and studied empirically (Stiglitz (1987), Laffont and Matoussi (1995), Tirole (1999)). Partly as a result, the terms trust, reciprocity, fairness, gift exchange and social capital now appear in the modeling and empirical study of principal agent relationships, the provision of public goods, and other standard economic applications, often in ways that underwrite mutually beneficial exchange consistent with Arrow's observation (Akerlof (1984), Benabou and Tirole (2006), Ellingsen and Johannesson (2008)).

Third, we may soon be able to isolate the neurological bases of the sometimes counterproductive effects of explicit incentives. Recent advances in brain imaging and other techniques have provided provisional identification of distinct brain regions whose activation is associated with the expression of social preferences (Greene, et al. (2001), Rilling, et al. (2002), Sanfey, et al. (2003)) and provided evidence suggesting that explicit incentives diminish activity in these social reward networks (Li, et al. (2008)).

Fourth, economists have increasingly turned to the study of cases in which preferences are not exogenous but rather are shaped by individuals' economic and other experiences, including their exposure to incentives of differing types (Becker (1996), Bowles (1998), Bisin and Verdier (2001), Bar-Gill and Fershtman (2005)).

Finally, beginning with the Lucas Critique of the exogenous beliefs assumption in macroeconomic policy, advances in the theory public policy have addressed cases in which incentives affect both beliefs and preferences and may thus may have unintended effects (Lucas (1976), Taylor (1987), Bowles (1989), Aaron (1994), Frey (1997), Bowles (2004), Bar-Gill and Fershtman (2005), Cervellati, Esteban and Kranich (2008)).

Here we extend the logic of the Lucas Critique to questions of framing, motivations, and social norms, in short, to preferences. To do this we modify the standard public economics and mechanism design assumption that taxes, subsidies, and other explicit incentives affect behavior only indirectly, that is by altering the economic costs and benefits of the targeted activities. In this conventional approach explicit incentives thus do not appear directly in the citizen's utility function and as a result, the behavioral effects of explicit economic incentives and social preferences are separable, the effects of each being independent of the levels of the other. We modify the citizen's utility function so that this separability property need not hold and as a result the two kinds of motivations may be either complements -- social preferences being heightened by incentives appealing to self-interest -- or substitutes, when explicit incentives are said to crowd out social preferences.

Our concern is not with the importance of other-regarding motives, but rather the plausibility of the separability assumption. Because it is so often implicit, it may help to identify what may be its first explicit statement by John Stuart Mill (1844): 97

> [Political economy] does not treat of the whole of man's nature...,... it is concerned with him solely as a being who desires to possess wealth, … it predicts only such ...phenomena ...as take place in consequence of the pursuit of wealth. It makes entire abstraction of every other human passion or motive

Incentives may have counter-intuitive and counter productive effects for reasons other than non-separability. Strong monetary incentives, for example, may over-motivate an agent leading to greater than the optimal level of arousal posited by the so called Yerkes-Dodson law. This appears to be the mechanism underlying the negative effects of high incentives found in three experiments by Ariely, et al. (2005). Similarly, if agents have an income target, monetary incentives may allow target attainment with less effort. Camerer, et al. (1997) suggest that this may explain why New York City taxi drivers work fewer hours when they are making more per hour. Neither of these mechanisms involves the non-separability of self regarding and other regarding preferences, which is the focus of this paper.

The experimental evidence for non-separability that we survey here would not be very interesting if they did not reflect real-life behavior. Testing for separability in natural settings is difficult, but generalizing directly from experiments even for phenomena much simpler than separability is often unwarranted. (Levitt and List (2007)). Consider, for example, the Dictator Game in which a one subject (the dictator) is assigned an endowment of money and asked to allocate some portion of it (including none) to a passive recipient. Typically more than 60% of subjects allocate a positive sum to the recipient and the average is about a fifth of the endowment. We would be sadly mistaken if we inferred from this that 60 percent of individuals would spontaneously transfer funds to anonymous passers by, or that the same subjects would offer a fifth of the bills in their wallet to a homeless person asking for help. Subjects who reported that they had never given to a charity allocated 60 percent of their endowment to a named charity in a lab experiment. (Benz and Meier)).

Most individuals are strongly influenced by the cues of appropriate behavior offered by the situation in which an action is taken (Ross and Nisbett (1991)), and there is no reason to think that experiments are immune to this context-dependent aspect of individual behavior. Validity concerns arise from four aspects of human behavioral experiments do not arise in most well-designed natural science experiments. First, experimental subjects typically know they are under an unknown researcher's microscope, inducing different behaviors than would

occur under total anonymity or under the scrutiny of neighbors, family or workmates. Second, interactions with other subjects are typically anonymous and without opportunities for ongoing face to face communication, unlike many social interactions. Third, subject pools may be quite different from the real-world populations of interest, in part due to the process of recruitment and self selection. Finally, many of the experiments that provide evidence for the salience of social preferences are deliberately structured as strategic interactions like the Ultimatum Game that give scope for ethical or other regarding behavior that may be absent in competitive markets and other important real world settings.

It is impossible to know whether these four aspects of behavioral experiments bias experimental results in ways relevant to the question of separability. For example, the fact that in most cases subjects are paid a "show up fee" to participate in an experiment might attract the more materially oriented who may have less salient other regarding preferences subject to crowding out. But the fact that many of the subject pools are students who have not faced the hard choices of making a living might work in the opposite direction. While warranting caution in generalizing the details of experimental behavior to the real world none of these validity concerns is sufficient to dismiss the experimental evidence that social preferences are important behavioral motivations and that the salience of these preferences may be affected by explicit incentives. This is especially the case when experiments identify motives that allow a consistent explanation of otherwise anomalous real world examples of crowding in or out.

In the next section we provide a taxonomy of cases where separability of social and conventional preferences does not hold. Because people often react to the mere presence of explicit incentives rather than their extent (Gneezy (2003)), we distinguish between categorical marginal effects. In the subsequent five sections we consider reasons why the separability may fail and provide experimental evidence about these five mechanisms. We conclude with some implications for policy and institutional design.

2. Incentives and social preferences as complements or substitutes

Consider an individual who may take an action that is costly to the actor and confers benefits on others. Taking the action may be encouraged by a subsidy or other explicit incentives (namely those that affect the expected material costs and benefits associated with the action.) Citizens also have "values" that may motivate taking pro-social actions, the term encompassing both ethical commitments and other-regarding preferences such as altruism. Where separability does not hold, the behavioral effects of these values may be influenced

(positively or negatively) by the use of explicit incentives. Assume that for a given individual the extent the action (denoted by a) and both incentives (s) and the intensity of values (λ_0) can be represented by a single number. Then we describe their interrelationships by the individual's choice of an action: $a^* = \mu(s, \lambda_0)$. Separability means that the effect of varying each of the arguments of μ is independent of the level of the other argument.

Non-separability may be either marginal (the effect of incentives on values depending continuously on the extent of the former) or categorical (the presence of incentives affecting values independently of their level) or a combination of the two. The presence of these discontinuous effects requires a more general definition of separability than the standard one, namely that the cross-partial derivative of $\mu(s, \lambda_0)$ be zero. Letting Δs and $\Delta\lambda_0$ represent arbitrary changes in incentives and values, separability implies that Δ^T, the effect on a^* of varying both s and λ_0 is equal to Δ^S, the sum of the effects of varying each separately where

(1) $\Delta^T \equiv \mu(s+\Delta s, \lambda_0+ \Delta\lambda_0) - \mu(s,\lambda_0)$ and

$\Delta^S \equiv \mu(s+\Delta s, \lambda_0) + \mu(s, \lambda_0+ \Delta\lambda_0) - 2\mu(s, \lambda_0)$

Where $\Delta^T > \Delta^S$ then incentives and social preferences are synergistic and are termed complements. Where the reverse is true the two arguments are substitutes (or are said to exhibit “negative synergy” or “crowding out”). Table 1 summarizes the relevant definitions and gives terms commonly used to refer to violations of separability.

Table 1. Separability and its violations.

$\Delta^T = \Delta^S$	Separability, additivity
$\Delta^T > \Delta^S$	Complementarity, synergy, super-modularity, crowding in
$\Delta^T < \Delta^S$	Substitutability, negative synergy, sub-modularity, crowding out

For concreteness, we study a single member of a community (indexed by j) who may contribute to a public project by taking an action a^j at a cost $g(a^j)$ that is increasing and convex in its argument. The output of the project is available to all and it varies positively and linearly with the sum of the n members’ contributions according to $\varphi(a^1 + ... + a^j + ... + a^n)$. The explicit incentive $s \geq 0$ is a payment to the individual that is proportional to the amount the individual contributes.

We express the individual's values as an addition to utility that is proportional to the level of contribution, and we make explicit the sources of non-separability as:

(2) $v^j = a^j \lambda_0(1+ 1\{s>0\}\lambda_1 + \lambda_2 s)$

where the indicator function $1\{s>0\} = 1$ if $s > 0$ and zero otherwise. In equation (2) as before $\lambda_0 \geq 0$ measures the intensity of values, λ_1 (which may be of either sign) measures the categorical effect of the presence of an incentive on values that is independent of the level of the incentive, and λ_2 (which also may be of either sign) measures the marginal (rather than categorical) effect of variations in s on values. The individual's utility is thus

(3) $u^j = \varphi(a^1 + \ldots + a^n) - g(a^j) + a^j(s + \lambda_0(1 + 1\{s>0\}\lambda_1 + \lambda_2 s))$

and the individual's utility maximizing contribution equates the marginal costs of contributing to the marginal benefits, or:

(4) $g'(a^*) = \varphi + s + \lambda_0(1 + 1\{s>0\}\lambda_1 + \lambda_2 s)$

Assuming that $g(a^j)$ is just $\frac{1}{2}(a^j)^2$ so as to permit a closed form expression for the individual's choice of contribution we have (ignoring the individual's superscript):

(5) $a^* = \mu(s, \lambda_0) = \varphi + s + \lambda_0(1 + 1\{s>0\}\lambda_1 + \lambda_2 s)$

and the effect of variations in s on the individual's actions is

(6) $\Delta a^* = \Delta s\,(1 + \lambda_0\lambda_2) + 1\{s=0\}\lambda_0\lambda_1$

Then using the fact that $1\{s=0\} + 1\{s>0\} = 1$ we have

(7) $\Delta^T = (\varphi + s + \Delta s + (\lambda_0 + \Delta\lambda_0)(1 + \lambda_1 + \lambda_2(s + \Delta s))) - (\varphi + s + \lambda_0(1 + 1\{s>0\}\lambda_1 + \lambda_2 s))$

$= \Delta s\,(1 + \lambda_0\lambda_2) + \Delta\lambda_0(1 + \lambda_1 + \lambda_2 s) + \Delta\lambda_0\Delta s\lambda_2 + 1\{s=0\}\lambda_0\lambda_1$

(8) $\Delta^S = (\varphi + s + \Delta s + \lambda_0(1 + \lambda_1 + \lambda_2(s + \Delta s))) + (\varphi + s + (\lambda_0 + \Delta\lambda_0)(1 + 1\{s>0\}\lambda_1 + \lambda_2 s))$

$- 2(\varphi + s + \lambda_0(1 + 1\{s>0\}\lambda_1 + \lambda_2 s))$

$= \Delta s\,(1 + \lambda_0\lambda_2) + \Delta\lambda_0(1 + 1\{s>0\}\lambda_1 + \lambda_2 s) + \lambda_0 1\{s=0\}\lambda_1$

Equality of Δ^T and Δ^S obtains if

(9) $\Delta^T - \Delta^S = \Delta\lambda_0(\Delta s\lambda_2 + 1\{s=0\}\lambda_1) = 0$

In (9) the first term in the parenthesis captures non-additivity due to marginal non-separability and the second non-additivity due to categorical non-separability. Figure 1 illustrates the two forms of non-separability.

Using (6) we say that a particular change in incentives Δs has crowded out values if $\Delta a^*/\Delta s < 1$, and conversely for the case of crowding in. Strong crowding out holds if $\Delta a^*/\Delta s < 0$. Note that crowding out does not require that the effect of the incentive be negative, only that it be less than would be the case if additivity held. This is illustrated in Figure 1 where the incentive has a positive effect on contributions in the presence of marginal (non-strong) crowding out and categorical crowding out with $s > s'$.

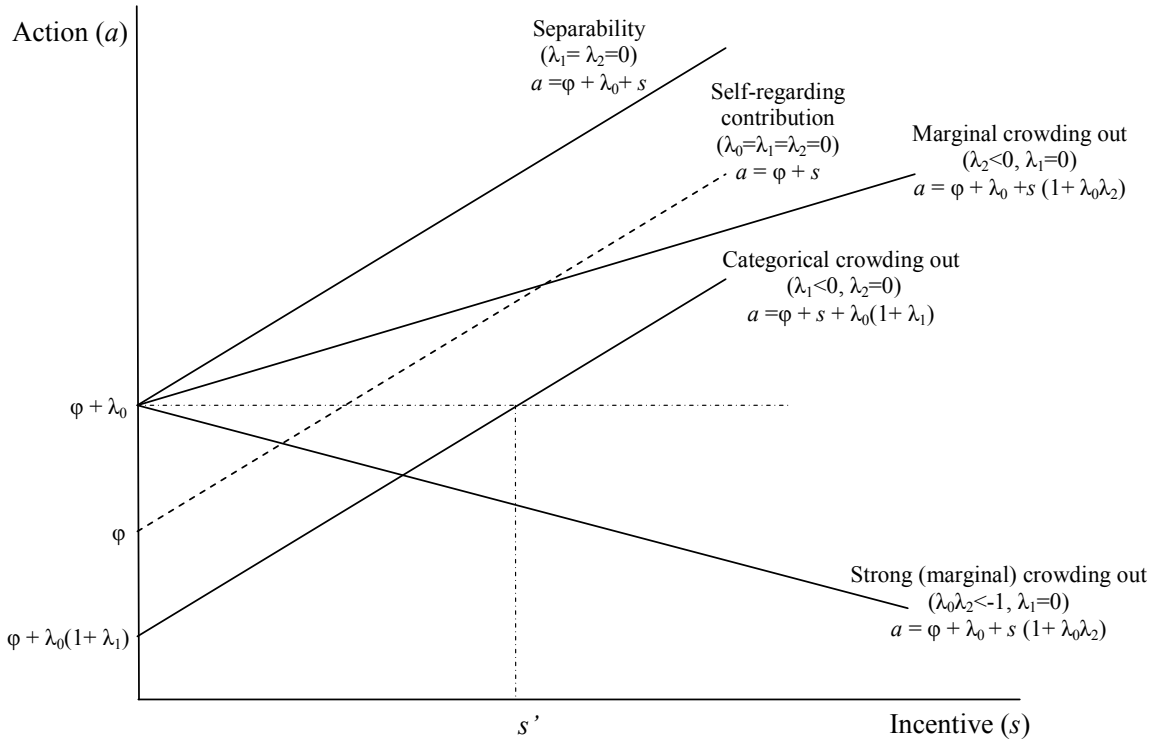

Figure 1. Citizen's contribution to the public good (a^*) under non-separability of incentives and values**.** Shown are examples of equation (5) under varying separability assumptions. Under separability (top line) categorical and marginal incentive effects are additive. Under strong crowding out the use of the incentive is counterproductive; this holds for all levels of s under the marginal crowding out function shown. Under categorical crowding out, incentives less than s' are also counterproductive in the sense that contributions are less than they would have been in the absence of incentives.

A recent experiment allows an estimate of both categorical and marginal crowding out. Irlenbusch and Ruchala (2008) implemented a public goods experiment in which the 192 German student subjects faced three conditions: no incentives to contribute and a bonus given to the highest contributing individual that was either high or low (details are in Table 2, results are shown in Figure 2). In the no-incentive case contributions averaged 48 percent above the Nash equilibrium that would have occurred if the participants had been motivated only by the material rewards of the public project (namely, 25). Contributions in the low-bonus case were not significantly different from the no-bonus treatment. In the high-bonus case significantly higher contributions occurred, but the amount contributed barely (and insignificantly) differed from that predicted for self-interested subjects.

In Figure 2 we use the observed behavior in the high and low bonus case to estimate a constant marginal effect of the bonus, finding that a unit increase in the bonus is associated with a 0.31 increase in contributions. This contrasts with the marginal effect of 0.42 that would have occurred under separability. Marginal crowding out thus affected a 26 percent reduction in the effect of the incentive. The estimated response to the incentive also gives us the level of categorical crowding out, namely the observed contributions (37.04) minus the

predicted contributions had an arbitrarily small incentive been in effect (the vertical intercept of the observed line in figure 2) or 34.56. The incentive thus categorically crowded out 21 percent of the effect of social preferences (measured by the excess in contribution levels above Nash equilibrium for self interested subjects) or 12.04. Categorical crowding out is also evident in three experiments by Heyman and Ariely (2004). For example reported willingness to help a stranger load a sofa into a van was much lower under a small money incentive than with no incentive at all, yet a moderate incentive increased the willingness to help (over the no incentive condition). Using these data as we did in the Irlenbusch and Ruchala study, we estimate that the mere presence of the incentive reduced the willingness to help by 27 percent (compared to the no incentive condition).

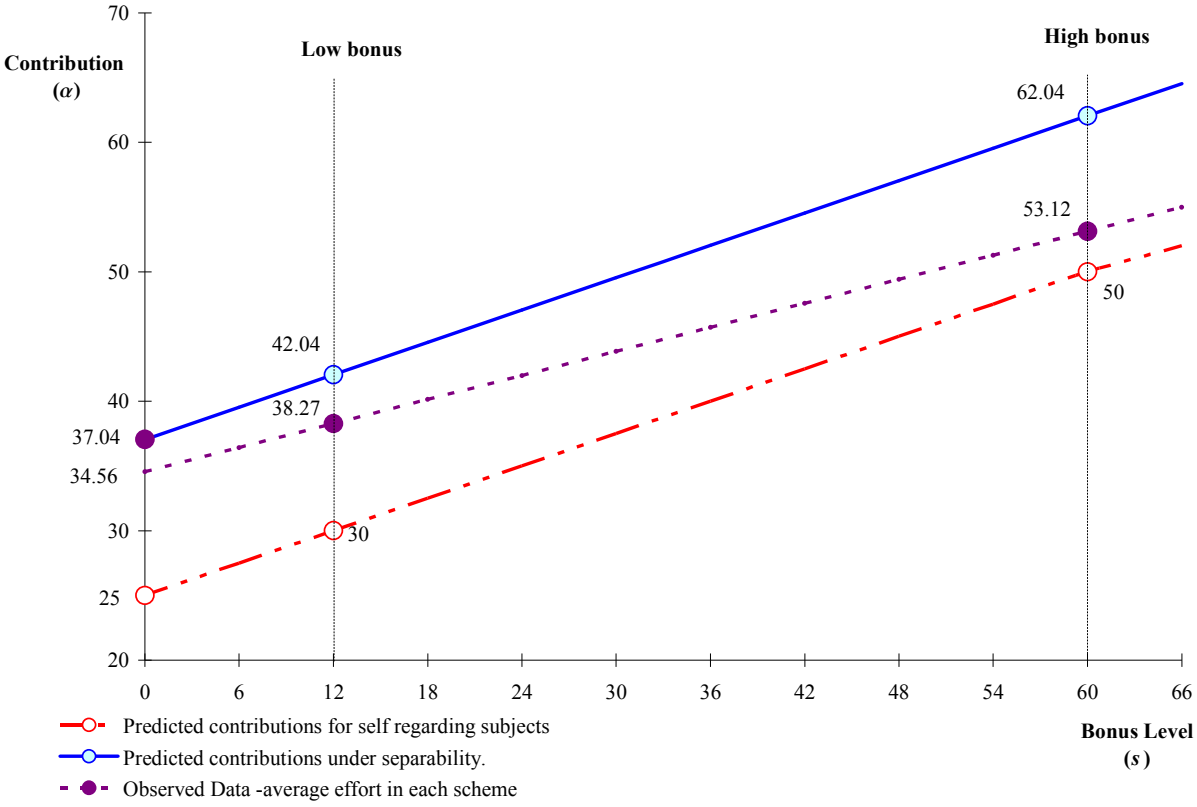

Figure 2. Categorical and marginal crowding out (from Irlenbusch and Ruchala (2008)). Source: see text. The experimental design is an adapted Voluntary Contribution Mechanism game comparing two team-based compensation schemes without and with a relative reward (or bonus) for the highest contributor in the team The bonus is self-funded (each member pays one-forth of the bonus). Each subject simultaneously decides an effort level from the interval [0, 120].

Many experiments provide evidence of strong but not weak forms of crowding out. The reason is that unlike the Irlenbusch and Ruchala (2008) study they do not establish the response to incentives that would be observed under separability and thus are able to detect only strong crowding out (based on the sign of the effect) and not weak (based on the size of the effect). A common misinterpretation of these experiments is that $\Delta a^*/\Delta s > 0$, as was

found in the Irlenbusch and Ruchala (2008) experiment, is evidence against crowding out (Rigdon (2009)).

What are the cognitive or affective effects of incentives that explain the categorical and marginal crowding out observed in this and other experiments? Experiments have not yet been designed to answer this question, so the inferences that we draw in the next 4 sections must be provisional.

The experimental methods that have become standard in economics include playing for real stakes, excluding deception, and making explicit use of game theoretic concepts to clarify the role of incentives. As experimental methods differ considerably across disciplines, and for reasons of space we limit the entries in the tables to experiments done by economists, although we are concerned we may be missing relevant literature provided by other disciplines. All of those studies include baselines to ascertain whether the incentives led to changes in subject's decisions. We refer to a number of important experiments done using other methods in the text.

3. Incentives provide information

Incentives are implemented for a purpose, and because the purpose is often evident to the target of the incentives, the target may also infer information about the person who designed the incentive, about his or her beliefs concerning the target, and the nature of the task to be done (Benabou and Tirole (2003), Fehr and Rockenbach (2003)). We will illustrate this incentives-as-signals mechanism by the contrasting positive response to fines imposed by peers in public goods experiments and negative response to fines imposed by experimental 'investors' and 'employers' in a principal agent experiment.

German students in the role of "investor" chose a costly action benefiting the other player, called the "trustee," who, knowing the investor's choice, could in turn provide a personally costly "back-transfer," returning a benefit to the investor (Fehr and Rockenbach (2003).) When the investor transferred money to the trustee, he also specified a desired level of the back-transfer. The experimenters implemented an incentive condition in which the investor had the option of declaring that he would impose a fine if the trustee's back-transfer were less than the desired amount. The investor could also decline the use of the fine, the choice of using or declining the fine option being taken prior to the trustee's decision. There was also a "trust" condition in which no such incentives were available to the investor.

Trustees reciprocated generous initial transfers by investors with greater back transfers. But the use of the fine reduced return transfers conditional on the investor's transfer, while

renouncing the use of the fine when it was available to the investor increased return transfers. Only one-third of the investors renounced the fine; their payoffs were 50 percent greater than the investors who threatened use of the fines.

The proximate neural causes of the negative impact of incentives in this case are suggested by evidence on the neural responses of the trustees in a Trust Game (Li, et al. (2008)). As in the experiment of Fehr and Rockenbach (2003) the investor's threat of sanctions negatively affected back transfers by trustees. To identify the proximate causes of this result, Li and his co authors used functional magnetic resonance imaging (fMRI) to compare the activation of distinct brain regions of trustees when faced with an investor who has threatened to sanction the trustee for insufficient back transfers and an investor who has not threatened a sanction. Activity in the Ventromedial Prefrontal Cortex (VMPFC) correlated with higher repayment by the Trustee. Threatened sanctions de-activated the VMPFC and other areas relating to the processing of social rewards, while activating the parietal cortex, an area thought to be associated with cost benefit analysis and other self interested optimizing. The interpretation by Li and his co authors is that the sanctions induced a "perception shift" favoring a more self interested response.

The interpretation suggested by Fehr and Rockenbach is that in the trust condition, or when the fine was renounced by the investor, a large initial transfer signaled that the investor trusted the trustee eliciting a positive reciprocal response. The threat of the fine, however, conveyed a different message and extinguished the trustee's reciprocity. This was especially the case when it appeared that the intent of the fine was to impose what the trustee considered to be an unfair outcome. Where the investor had announced modest levels of desired returns such that the investor and the trustee would both share in the benefits, the use of the fines reduced back transfers by an insignificant 8 percent. But where the announced desired back-transfer would have allowed the investor to capture most of the benefits had the trustee complied, the reduction in back transfers was 38 percent.

The fact that in this case incentives reveal that the principal is untrusting or self-aggrandizing helps explain the contrasting effect of incentives imposed by peers who do not stand to benefit personally. An example is the public goods experiment in which fellow group members have the opportunity to reduce their own payoffs in order to punish (reduce the payoffs of) others in their group once each member's contributions are revealed, and in which group membership is shuffled so that a punisher could not benefit from the target's response in subsequent periods. In this setting there is a strong positive response by low contributors (Fehr and Gachter (2000), Fehr and Gaechter (2002a)). The most plausible explanation of the

effectiveness of incentives in this case is that when punished, those who have contributed less than others interpret the punishment as a signal of public-spirited social disapproval and feel shame, which they redress by contributing more subsequently. If those punished experience anger instead of shame (if they have contributed more than others, for example), they may subsequently contribute less, and costly retaliatory punishment escalations result (Bowles and Gintis (2006), Hopfensitz and Reuben (2006), Carpenter, et al. (2009).) Table 2 summarizes experiments in which this incentives-as-signals effect appears to have been at work (in some cases along with other mechanisms, to which we now turn).

4. Incentives may suggest appropriate behavior

In most situations people look for clues of appropriate behavior (Tversky and Kahneman (1981), Ross and Nisbett (1991), Salant and Rubinstein (2007)) and incentives often provide them. In Table 3 we survey experiments in which this appears to have been the case.

Schotter, Weiss and Zapater (1996) found that competition for "survival" among subjects playing the Ultimatum Game induced lower offers and fewer rejections of low offers, commenting: "...the competition inherent in markets...offers justifications for actions that, in isolation, would be unjustifiable." Hoffman, et al. (1994) were the first to discover the power of names: generosity and fair-minded behavior were diminished by simply re-labeling an Ultimatum game the "Exchange Game" and re-labeling proposers and responders "sellers" and "buyers". The power of names has been confirmed in many experiments since (cited in Rege and Telle (2004), Zhong, Loewenstein and Murnighan (2007), Ellingsen, et al. (2008)). But literally naming the game is not necessary. Incentives alone may provide powerful frames for the decision maker. For example, a weak sanctioning system may result in less cooperation than no sanctioning system; sanctions affect the type of decision people perceive they are making, prompting them to see it as a business rather than an ethical decision. Subjects evaluate sanction treatment (both weak and strong) as 'business' rather than 'ethical' (Tenbrunsel and Messick (1999))

This appears to have been the case in an experiment with rural people in Colombia. Experimental subjects whose livelihoods depend on easily-depleted forest resources were asked to individually and anonymously choose how much to withdraw from a mutually beneficial common pool called 'the forest' (Cardenas, Stranlund and Willis (2000)). Groups of subjects played 8 rounds of this game without communication, withdrawing on average amounts that were about midway between the individually self-interested and the group-

beneficial levels. Their substantial deviation from the individually selfish level is a measure of the subjects' other-regarding or ethical values. In subsequent play for some groups, face-to-face cheap talk communication was allowed. Groups in this “communication” treatment improved their performance, extracting somewhat less from the 'forest', thereby deviating more from self interest, and gaining higher benefits as a result.

The other treatment precluded communication but simulated a "government regulation". Withdrawals were not to exceed the announced group-optimum level, and subjects would be monitored and fined for over-exploitation. The regulation reduced the level of withdrawal that would be chosen by an entirely selfish individual, but the expected fines were such that some overexploitation of the common pool remained the payoff maximizer’s optimal choice. In this “regulation” treatment, subjects initially responded by restricting their withdrawals to close to the group optimum. But after three rounds their behavior increasingly conformed to self-interest, and for the last three rounds their choices were almost entirely self-interested, sacrificing only one-quarter as much individual payoff to protect the 'forest' as they did in the final three rounds of prior to the imposition of the incentive.

The most plausible explanation is that the fine, while insufficient to enforce the social optimum, displaced the subjects' ethical predispositions that in the earlier rounds had induced them to withdraw much less than would maximize their own payoffs. We do not have direct evidence for this explanation because the social preferences accounting for the more-than-selfish levels of contributions prior to the imposition of the fine were not measured. There are cases, however, in which the reduction in the salience of ethical reasoning induced by the presence of incentives can be identified. An example follows.

A large team of anthropologists and economists implemented both dictator and third party punishment games in 15 societies ranging from Amazonian, Arctic and African hunter gatherers to manufacturing workers in Accra, Ghana and U.S undergraduates (Barr, et al. (2009).) In the dictator game an experimental subject is assigned a sum of money and asked to award some all or none of it to an entirely passive respondent. The third party punishment game is a dictator game with an active onlooker (the third party) observes the dictator’s offer to the passive respondent. If the third party deems the dictator’s offer worthy of punishment he or she may then pay to impose a fine on the dictator. Though one would expect that the dictators would adjust their offers upwards to avoid being fined, fining was common; it occurred in 30% of the interactions across the study sites.

Surprisingly, in only two of the 15 populations were the offers significantly higher in the third party punishment game than in the dictator game, and in four of the populations the offers were significantly (and in some cases substantially) lower. In Accra, for example, where 41 percent of the dictator's offers resulted in fines, the amounts offered were 30 per cent lower (t = -6.8) in the third party punishment game than in the dictator game The incentives provided by the fine did not induce higher offers, but rather had the opposite effect. (The fact that for two groups there was a significant positive effect of the fine option indicates that the incentive had some effect, but does not preclude crowding out.)

5. Incentives may compromise intrinsic motives and self-determination

A rich experimental and theoretical literature in psychology has explored the crowding out of intrinsic motives (Deci and Ryan (1985), Deci, Koestner and Ryan (1999), Cameron, Banko and Pierce (2001).) Recent experiments by economists surveyed in Table 4 as well as non-experimental studies in economics (surveyed in Frey and Jegen (2001)) are consistent with this view. The underlying psychological mechanism appears to be a fundamental desire for "feelings of competence and self-determination that are associated with intrinsically motivated behavior". According to this interpretation where people derive pleasure from an action *per se* in the absence of other rewards, the introduction of explicit incentives may 'over-justify' the activity and reduce the individual's sense of autonomy.

Consistent with this "self-determination" model, Falk and Kosfeld (2006) explored the idea that 'control aversion' may be a reason why incentives degrade performance. Experimental agents in a role similar to an employee chose a level of 'production' that was costly to them and beneficial to the principal (the employer). The agent's choice effectively determined the distribution of gains between the two, with the agent's maximum payoff occurring if he produced nothing. Before the agent's decision, the principal could elect to leave the choice of the level of production completely to the agent's discretion, or impose a lower bound on the agent's production (three bounds were varied by the experimenter across treatments, the principal's choice was whether or not to impose it.) The principal could infer that a self-regarding agent would perform at the lower bound and thus imposition of the bound would maximize the principal's payoffs.

But in the experiment, agents chose a lower level of production when the principal imposed the bound. Apparently anticipating this response, less than a third of the principals opted for its imposition in the moderate or low bound treatments. This minority of "untrusting" principals earned on average half of the profits of those who did not seek to

control the agents' choice in the low bound treatment, and a third less in the intermediate bound condition. In post-play interviews, most agents agreed with the statement that the imposition of the lower bound was a signal of distrust.

Control aversion and the desire for self-determination are not the only effects of the principal's seeking to bind the agent. As anticipated by our discussion of the information content of incentives above, the imposition of the minimum in this experiment gave the agents remarkably accurate information about the principals' beliefs concerning the agents: those who imposed the bound had substantially lower expectations of the agents. Their consequent attempt to control the agents' choices induced over half of the agents (in all three treatments) to contribute minimally, thereby affirming the principals' pessimism. Depending on the distribution of principal's priors about the agents, a population with preferences similar to these experimental subjects could support both trusting and untrusting (Pareto-inefficient) equilibria.

6. Incentives alter the environment in which new preferences are learned

Incentives may also induce long-term change in motivations because they affect key aspects of how we acquire our motivations, including both the range of alternative preferences to which one is exposed and the economic rewards and social status of those with preferences different from one's own (Bisin and Verdier (2001), Bowles (2004), Bar-Gill and Fershtman (2005)). For example, suppose the relevant incentives allow the selfish to exploit the civic-minded, then if the learning process is payoff-monotonic the civic-minded are likely to be eliminated. Other effects are less obvious: a competitive market with complete contracts leaves little scope for acting on ethical, reciprocal or generous preferences, even among those so inclined (Sobel (2007)). If preference change is closely related to exposure to alternative models (Zajonc (1968)), then this idealized market environment would provide little basis for the proliferation of non-self-interested preferences.

Experiments of just a few hours duration are unlikely to uncover the causal mechanisms at work. This is because adopting new preferences is often a slow process more akin to acquiring an accent than to choosing an action in a game. The developmental processes involved typically include population-level effects such as conformism, schooling, religious instruction and other forms of socialization that are not readily captured in experiments. However, historical, anthropological, social psychological and other data (surveyed in Bowles (1998)) show that economic structures affect parental child rearing values, personality traits rewarded by higher grades in school, and other developmental

influences. Experiments in 15 small scale societies with extraordinarily varied economic structures ranging from farming to hunting and gathering revealed a strong association between the nature of the very diverse economic tasks required to secure a livelihood in a society and its members' behavior in the Ultimatum game

Despite the limitations of experiments for the investigation of preference change, we survey in table 5 a number of experiments that have documented durable learning effects. In many cases incentives induced more self-interested behavior, even after they were withdrawn. In the public goods experiment designed by Falkinger, et al. (2000) an incentive system induced subjects to contribute almost exactly the amount predicted for a own-material-payoff-maximizing individual. In the absence of the incentive Falkinger's subjects contributed significantly more than would have been optimal for a payoff maximizing individual. But subjects who had previously experienced the incentive contributed 26 per cent less than those who had never experienced it.

7. Incentives and social preferences as complements

Crowding in may also occur. We have already seen that fines imposed on free riders by altruistic peers in a public goods game induce higher levels of contribution in subsequent rounds of play. Of course individuals might have simply best-responded to the anticipated loss in payoffs associated with low contributions; but more than this appears to be at work here. In the public goods experiment designed by Carpenter, et al. (2009) contributing nothing the public good remained a best response for a self interested individual even when punishment of miscreants was allowed (the observed punishment, while substantial was insufficient or offset the cost of contributing). Nonetheless individuals responded positively to having been punished for low contributions in previous rounds, consistent with the hypothesis that punishment heightened the salience of shame or other social emotions. Consistent with this interpretation, purely verbal messages of disapproval have a substantial positive effect on free riders' subsequent contributions (Barr (2001), Masclet, et al. (2003).)

Apparently in this case the fines evoke shame and lead the individual to seek to avoid this unpleasant emotion in the future. But other mechanisms are at work: social norms support the observance of traffic regulations, but these may unravel in the absence of state-imposed sanctions on flagrant violators. The rule of law and other institutional designs that limit the more extreme forms of anti-social behavior and facilitate mutually beneficial interactions on a large scale may enhance the salience of social preferences by assuring people that those who conform to moral norms will not be exploited by their self-interested fellow citizens. This

phenomenon may have been at work among the Hokkaido University subjects who cooperated more in a public goods experiment when assured that others who did not cooperate would be punished (Shinada and Yamagishi (2007)) despite the fact that this had no effect on their own material incentives. They apparently wanted to be cooperative but wished even more to avoid being the sucker who is exploited by defectors. Market incentives may also favor the endogenous evolution of social preferences. In experiments in 15 small-scale societies in Africa, Asia and Latin America (Henrich, et al. (2005)), the experience of mutually beneficial exchanges with strangers may explain why, in anonymous experimental settings, individuals from the more market integrated societies were also the most fair-minded.

A distinct mechanism underlying crowding in was apparently at work in a public goods experiment by Galbiati and Vertova (2008b). They found that the effect of a stated (non-binding) obligation to contribute a certain amount was greater when it was combined with a weak monetary incentive than when no incentives were offered. The monetary incentives had no effect on behavior in the absence of the stated obligation. The authors' interpretation is that the explicit incentives enhanced the salience of the stated obligation.

8. Conclusions: Puzzles and lessons

While each experiment may bear diverse and even competing interpretations, in light of these data it would nonetheless be difficult to sustain the standard economics separability assumption. The most plausible explanation is based on the fact that when people engage in trade, produce goods and services, save and invest they are not only attempting to get things, they are also trying to be someone, both in their own eyes and in the eyes of others. This may explain why incentives for settlement of conflicts may fail. Representative samples of Jewish West Bank settlers in 2005, Palestinian refugees in 2005, and Palestinian students in 2006 were asked how angry and disgusted they would feel or how supportive to violence they might be should their political leaders were to compromise on contested issues between the groups. Those who regarded their group's claims as reflecting sacred values (about half in each of the three groups) expressed far greater anger, disgust and support for violence should the compromise be accompanied by a monetary compensation for their own group than if no compensation was offered. (Frey and Stutzer (2006), Ginges, et al. (2007).)

John Stuart Mill and economists since have recognized that the purposes of individual economic action are constitutive as well as acquisitive; what many appear to have missed is that incentives addressed to our acquisitive side interact as either complements or substitutes

with our constitutive projects. Jeremy Bentham's *Introduction to the Principles of Morals and Legislation* (1789), arguably the first text in what we now call public economics, understood the constitutive side of action and its importance for the design of public policy:

> A punishment may be said to be calculated to answer the purpose of a moral lesson, when by reason of the ignominy it stamps upon the offence, it is calculated to inspire the public with sentiments of aversion towards those pernicious habits and dispositions with which the offence appears to be connected; and thereby to inculcate the opposite beneficial habits and dispositions (Bentham (1789): p.26.)

The fact that punishments are "moral lessons" as well as incentives may help resolve one of the puzzles in the literature we have just surveyed. In a widely cited natural experiment, the imposition of fines on parents arriving late to pick up their children a day care centers in Haifa resulted in a doubling of the number of tardy pickups (Gneezy and Rustichini (2000a)). But the small tax on plastic grocery bags enacted in Ireland in 2002 had a the opposite effect: it resulted in a 94 percent decline in their use and appeared to crowd in pro-social preferences (Rosenthal (2008).).

The contrast is instructive. In the Haifa case, the experimenters (respecting standard experimental protocols) provided the parents no justification for the introduction of the fine, whose occasional lateness could have occurred for reasons beyond the parents' control rather than disregard of the inconvenience it caused. Moreover lateness was not very common and hence was not widely broadcast to the other parents. By contrast, the introduction of the Irish plastic bag tax was preceded by an effective publicity campaign and the use of the bags was the result of a simple choice made in a highly public condition. In the Irish case, as in the experiment by Galbiati and Vertova (2008b), the monetary incentive was introduced jointly with a message of explicit social obligation, and it apparently served as a reminder of the larger social costs of the use and disposition of the bags. This contrast, along with the fact mentioned above that fines imposed by peers in public goods games have positive effects while fines imposed by principals on agents often backfire, makes it clear that fines and other monetary incentives *per se* are not the cause of crowding out. Rather what is critical is the meaning of the fines as conveyed by the social relationships among the actors, the information the fine provides, and the pre-existing normative frameworks of the actors.

Another lesson for mechanism design is that in implementing public policy or private systems of incentives, the designer must consider the response of individuals' motivations to the instruments under consideration and take the predicted policy outcome to be the resulting joint equilibrium of preferences and economic allocations. Perhaps surprisingly, the citizen-utility-maximizing sophisticated planner cognizant of this motivational version of the Lucas

critique may make either greater or lesser use of explicit incentives when crowding out occurs (Fershtman and Heifetz (2006), Heifetz, Segev and Talley (2007), Bowles and Hwang (2008).)

Tables

Note: The bold entries **F, S, E** and **C** indicate that the experiment in question could also have been included in tables 3 (**F**raming) 4 (**S**elf-determination) 5 (**E**ndogenous preferences) or 6 (**C**omplementary relations between incentives and social preferences). In those tables **I** indicate that the experiment could have been included in this table (**I**nformation). All the papers but those marked with an * are published or forthcoming in a publication.

Table 2. Incentives provide information (I)

Citation	Subjects (number)	Games or activities	Institutional environments (treatments)	Results relevant to separability	Comment (quotes are from the cited paper)
Ariely, Bracha and Meier (2009) [01]	U.S. students (161)	Charity giving based on task performance	• Two external forms of enforcement: With monetary compensation or without; • Donation choices are public or private • Different frames: "good" and "bad" charitable causes	In the public treatment subjects exert more effort for a good cause and gifts are substantially lower in the incentive treatment. Monetary incentives increase effort in the private treatment.	The signaling value of giving is compromised by incentives. "image motivation is crowded out by monetary incentives [that are] more likely to be counterproductive for public pro-social activities than for private ones." (p.1) Categorical crowding out. See Tenbrunsel and Messick (1999), Mulder, et al. (2006) and Upton (1974)
Irlenbusch and Ruchala (2008) [02]	German Students (192)	Public goods game	• Two external forms of enforcement: Team-based compensation with and without a relative reward for the highest contributor in the team • The relative reward is a low or a high bonus • Pure Individual bonus without team-based compensation	High (but not low) bonuses increase average effort, and joint surplus increases significantly only if the bonus is high, but decreases over time. Only with the purely team-based compensation (no individual incentives) do agents contribute more than self interest would motivate. Pure tournament incentives induce effort levels below the selfish Nash equilibrium prediction.	Both categorical and marginal crowding out occur. The tournament structure reduces voluntary cooperation. **F**
Borges and Irlenbusch (2007) [03]	German Students (179)	Buyer- seller	• Three rights of withdrawal: none, voluntary offer of a right of withdrawal (with a return cost for the seller) and imposed. • The right of withdrawal when imposed has a return cost for the buyer or not	When sellers voluntarily offer a withdrawal right, buyers make order decisions that are less harmful for the seller than if the withdrawal right is imposed on sellers exogenously.	"Buyers are more inclined to behave fairly towards the sellers if they have granted the withdrawal right voluntarily than if it is constituted by law". (p. 17) [because it is] "perceived ...as a generous act and they might feel inclined to reciprocate by not exploiting the seller. …". (p. 12) **F**
Dickenson and Villeval (2007) [04]	French students (182)	Gift-exchange game with a computer task	• Stranger or Partner with communication • Employer payoffs dependent on employee effort (variable) or not.	In the variable-partner treatment (but not in the others) less monitoring induce substantially higher performance. Consistent with Frey (1993)	While intrinsic motivation is evident in subject behaviors, in the Partner relationship the effect of more monitoring appears to be a reciprocity-based negative response to the principal's lack of trust or intent to benefit at the agent's expense. **F, S**

Table 2 continued…

Citation	Subjects (number)	Games or activities	Institutional environments (treatments)	Results relevant to separability	Comment (quotes are from the cited paper)
Stanca, Bruni and Corazzini (2007) * [05]	Italian students (96)	Gift-exchange game	• In the first move, Information (player 1 knows there is a second move) or No Information (player 1 does not know there is a second move and hence thinks the game is a Dictator game)	Second movers' amounts returned are more correlated with the first mover's amounts sent in the No Information treatment.	Reciprocity is stronger in response to actions that are perceived as driven by intrinsic motivation, than in response to actions that are perceived as extrinsically motivated. **F**
Tyran and Feld (2006) [06]	Swiss students (102)	Public goods game	• Levels of sanctions: none, mild and severe • Enforcement: external (i.e. experimenter-imposed) or self-imposed (by referendum)	Exogenously imposed mild law does not significantly affect average contributions to the public good. Compliance is much improved if mild law is endogenously chosen.	If the enforcement is self-imposed it does not indicate hostile intent and also induces expectations of others' cooperation (people tend to comply with the law if they expect many others to do so). If mild law is rejected in the referendum, compliance tends to be lower than without the law. **F**
Fehr and List (2004) [07]	Costa Rican CEOs (126) & students (76)	Trust game	• Optional punishment as an incentive contract (i.e. a fine if less than the desired back-transfer amount is returned)	CEO principals trust more and are more trustworthy than students and as a result they achieve allocations closer to the maximum surplus that could be generated by the two parties. Joint surplus is highest when the punishment option is available and not used and lowest if the punishment option is used.	Key to performance: "the psychological message…conveyed by incentives – whether ... kind or hostile..." (p. 745). See Fehr and Rockenbach (2003) [08]
Fehr and Rockenbach (2003) [08]	German students (238)	Trust game	• Optional punishment as a incentive contract (i.e. a fine if less than the desired back-transfer amount is returned)	Trustee's back-transfers are lower when investors impose fines. Not using the punishment option when it is available results in larger back transfers and a larger joint surplus.	Explicit incentives undermine altruistic cooperation and reciprocity; forgoing the punishment option is a signal of good will and trust. See Fehr and List (2004) [07] Negative effects of use of the punishment option are greater when the investor demands a larger share of the joint surplus. Categorical crowding out when the investor chooses the fine. **F**
Fehr and Gaechter (2002b) * [09]	Swiss students (182)	Gift-exchange game	• Three external forms of enforcement: A Trust (pure fixed wage) contract, a price deduction (i.e., fine) contract, and bonus incentive contract	Incentives reduce agent's effort. If the incentive is framed as a price deduction the effort reduction is greater than where the incentive is framed as a bonus. Incentives reduce total surplus, increase principal's profits.	Effects of incentives are due to the perceived fairness, kindness and hostility of the principal's action. **F, S**

Table 3. Incentives may suggest appropriate behavior (F)

Citation	Subjects (number)	Games or activities	Institutional environments (treatments)	Results relevant to separability	Comment (quotes are from the cited paper)
Ellingsen, et al. (2008) * [10]	Swedish students (668)	Prisoners' dilemma game	• Two labels: Community Game and the Stock Market Game • Two types of informed interactions: human - human and human – computer (opponent's choice of action is made by a computer that is programmed to play with the same frequency as do subjects in the human - human treatment)	Cooperation is higher with the Community Game label than the Stock Market game label in the human-human interaction. There is no such effect in the human –computer interaction: there is no labeling effect when subjects play against an opponent who is unaware of the game, although the opponent's action is guaranteed to be statistically identical to the actions of an informed opponent.	Cooperative language does not suffice to increase cooperation. People respond to labels because the label affects how others interpret their behavior, which in turn determines their image. "people's behavior is constantly sensitive to whom they are interacting with and what these opponents will do and think" (p. 8). See Zhong, et al. (2007), Ross and Samuels (1993), Ross and Ward (1996)
Galbiati and Vertova (2008b) [11]	Italian students (210)	Public goods game (and a lottery game)	• Different levels of the obligated contribution (zero, low and high) with a low level of explicit incentives (i.e. a probability of monitoring and a probabilistic penalty or reward)	When the obligated contribution required is high, cooperation is significantly higher than in presence of low or null obligation, despite the material incentives being identical in these cases.	Obligations (i.e. what formal rules ask people to do) affect behavior independently of economic incentives. **I**
Li, et al. (2008) * [12]	US citizens (104)	Trust game	• Optional punishment as an incentive contract (i.e. a monetary sanction if less than the desired back-transfer amount is returned)	Trustees reciprocate relatively less when facing sanction threats, and the presence of sanctions significantly reduces trustee's brain activities involved in social reward valuation (VMPFC, LOFC, and amygdala), while simultaneously significantly increasing activities in parietal cortex previously implicated in rational decision making.	Monetary sanctions "encourage activity within neural networks associated with self-interested economic decision making while simultaneously mitigating activity in networks implicated in social reward evaluation and processing" (p. 3) **I**
Bohnet and Baytelman (2007) [13]	Senior executives in U.S. (353)	Trust game and a Dictator game (for trustors the transfer is tripled and for trustees the transfer does not change)	• No communication, face-to-face pre-play communication or post-play communication • Two external forms of enforcement (Post-play monetary punishment or not) • Stranger and Partner	Repetition and communication increase amount sent and returned; the option of punishment for low offers reduces offers of other- regarding trustees.	"The availability of punishment destroys intrinsic trust and lowers people's willingness to reward trust" (p.1) **I**

Table 3 continued…

Citation	Subjects (number)	Games or activities	Institutional environments (treatments)	Results relevant to separability	Comment (quotes are from the cited paper)
Houser, et al. (2007) [14]	U.S. students (532)	Gift Exchange game	• Two forms of enforcement (Punishment as an incentive contract (i.e. a fine)) • Intention treatment: Punishment is assigned randomly or imposed by investors	When back-transfer requests are high in relation to the sanction's size, regardless of whether the request is fair and regardless of whether punishment is intentional, punishment incentives have detrimental effects on the amount returned.	"Subjects interpret punishment as the price for self-interested behavior and the price, regardless of whether it was intentionally imposed, is an excuse for selfishness" (p.15) Categorical crowding out when the investor chooses the fine. See Fehr and Rockenbach (2003) [08] and Mulder, et al. (2006) **I**
Fischbacher, Fong and Fehr (2005) * [15]	Swiss students (238)	Ultimatum game	• Buyer competition (one, two or five Responders) • Seller competition (one or two Proposers)	Buyer competition reduces mean accepted offers and buyers' willingness to reject.	Buyer competition makes punishment of 'unfair' offers less certain (buyers' expectations about other buyers' acceptance is less certain). Competition among responders appears frame the interaction market-like. **S**
Cardenas (2004) [16]	Colombian users of rural ecosystems (265)	Common pool resource game	• Different levels of external enforcement (weak and strong) with announcement of socially optimal extraction level and without communication • Communication without fines and announcement.	Deviation from self interested behavior is much greater under communication (no fine) than under either high or low fines without communication. The behavioral effect of high (rather than low) fines is less than 6 percent of the predicted effect assuming self interest.	"A significant fraction of individuals were more responsive to the norm of cooperation that was proposed externally [the announced optimal level] than to the expected cost of the regulation." (p. 238). **C**
Heyman and Ariely (2004) [17]	240 US students (150+90)	A Computer Task and a Puzzle Task	• Different forms of compensation (cash, candy or cash in terms of candy) • Different levels of monetary compensation (none, low, medium)	Effort in both the cash and the candy conditions increases when the compensation level increases from low to medium. Effort in the no-compensation treatment is higher than the low-compensation condition for both the cash and the cash in terms of candy conditions and is not different from low-compensation in the candy condition. Performance from no-compensation to low-compensation conditions decreases only with monetary exchange mechanisms.	The level and form of compensation affect performance. "Monetary compensation may act as a strong signal invoking norms of money markets instead of social-market relations" (p. 6) Monetary incentives influence the ways in which tasks are framed and the motivation to engage in them. The type of market in which the exchange takes place influences the relationship between reward and motivation. **I**

Table 3 continued…

Citation	Subjects (number)	Games or activities	Institutional environments (treatments)	Results relevant to separability	Comment (quotes are from the cited paper)
Cardenas, et al. (2000) [18]	Colombian forest area dwellers (112)	Common pool resource game	• External enforcement device with a weak inspection and a fine • Communication	Fines induce more self-interested behavior and common pool over-exploitation. Socially optimal deviations from the selfish Nash equilibrium behavior (and the implied foregone payoffs by subjects) are least under the fines.	Weakly (exogenously) enforced fines diminish socially motivated behavior. Fine appear to have induced a shift from moral to self interested frame. See Tenbrunsel and Messick (1999)
Schotter, et al. (1996) [19]	U.S. students (247)	Ultimatum and Dictator games	• Survival treatment (two-stage): subjects with higher payoffs "survive" to proceed to stage 2. • Non survival treatment (one stage): the proposer is randomly assigned • Contextual framing: a simultaneous move-normal or a sequential extensive form game	Competitive threats to survival induce lower offers, and in the UG fewer rejections of low offers.	The context affects behavior: 'earning' right to be the first mover or threat to survival induces proposers to behave in a more self-regarding manner. "…the competition inherent in markets and the need to survive offers justifications for actions that, in isolation, would be unjustifiable". (p.38) **S**
Hoffman, et al. (1994) [20]	U.S. students (270)	Ultimatum game; Dictator game	• Roles are assigned by contest (the right to be the Proposer is 'earned') or randomly assigned). • Different frame: "Exchange" game (between a "seller" and a "buyer") or no frame • Anonymity: Double blind or not	Offers are lower and fewer low offers are rejected in an exchange context or when the proposer earns the right to his role. Proposers accurately gauge willingness of responders to accept lower offers. Dictators send lower amounts in double blind.	Institutional cues affect behavior: with property rights (i.e. legitimate 'earning' right to be proposer), a market framing or total anonymity, proposers and responders are more self-regarding. **S**
Mellstrom and Johannesson (2008) [21]	Swedish students (262)	Subjects are offered to carry out the health exam to become blood donors	• With and without a monetary compensation for becoming blood donors • To choose between a monetary compensation and donating the same amount to charity	The incentive reduces the supply of female prospective blood donors from 52% to 30% among women. No effect among men. Allowing individuals to donate the payment to charity counteracts the negative effect of the monetary compensation.	The monetary incentive may make it more difficult to signal social preferences, diminishing the signaling value of contributing. Charity option facilitates signaling. Over-justification appears also to be involved. See Upton (1974). **I**

Table 4. Incentives may compromise intrinsic motives and self-determination (S)

Citation	Subjects (number)	Games or activities	Institutional environments (treatments)	Results relevant to separability	Comment (quotes are from the cited paper)
Fehr, et al. (2007) [22]	German students (130)	Gift-exchange game	• Three internal forms of enforcement: The principal can choose to rely on - a trust (pure fixed wage) contract, or a price deduction (i.e., fine) contract - a trust, a fine or an unenforceable bonus contract • Different frames: employer-employee or buyer-seller	Bonus contracts yield higher joint surplus than the fine contract; principals converge towards the bonus contract. Trust contracts yield lower joint surplus than incentive contracts and bonus contracts. Agents spend more effort under a bonus contract than under a fine contract. The results are the same independently of the framing.	Effectiveness of incentive contracts may depend on the agent's other regarding preferences: With fair-minded players, incomplete contracts that rely on fairness as an enforcement device (i.e., bonus) provide powerful incentives, superior to explicit incentive contracts. **I**
Fehr and Schmidt (2007) [23]	German Students (three sessions, each with 22-24 subjects)	Gift-exchange game	Two internal forms of enforcement: The principal can choose to rely on - an announced unenforceable bonus contract - a combination of the bonus contract with a fine.	Most principals do not use the fine. The joint surplus under the pure bonus contract is 20 percent greater than under the combined contract. Wages are 54 percent higher in the pure bonus contract. Profits are not significantly different in the two contracts.	"Explicit and implicit incentives are substitutes rather than complements" (p. 3). Agents perceive that principals who are less fair are more likely to choose a combined contract and less likely to pay the announced bonus. The effect of effort on the bonus paid is twice as great in the pure bonus case. **I**
Falk and Kosfeld (2006) [24]	Swiss students (804)	Principal–agent game	• Different levels of control for the minimum level of performance (low, medium, and high) • The levels of control are external (medium) or imposed by principals • A gift exchange game: the principal decides whether to control the agent and also determines agent's wage	Most agents perform minimally as a response to the principals' controlling decision. Majority of the principals anticipates this and decide not to control, earning higher profits as a result.	Control and explicit incentives are signals of distrust and low expectations, diminishes agents' reciprocity and good will towards the principal. Categorical crowding out. **I**
Gneezy (2003) [25]	U.S students (400)	Proposer-responder game	• The responder has three forms of enforcement (a punishment at a given cost, a reward at a given cost and nothing) • Different levels of the responder's enforcement (weak, strong)	Non-monotonic effects of explicit incentives (fines and rewards) on performance (a W -shaped function). Offers are highest with large incentives (fine and reward), and lowest with small incentives. The no incentive case, when proposers simply dictate allocation, is intermediate.	Extrinsic incentives undermine intrinsic motivation: a small fine or reward changes the mode of behavior from "moral" to "strategic". See Gneezy and Rustichini (2000a), Gneezy and Rustichini (2000b) and Mulder, et al. (2006) [27] [28] [35]. Categorical crowding out. **F**

Table 4 continued…

Citation	Subjects (number)	Games or activities	Institutional environments (treatments)	Results relevant to separability	Comment (quotes are from the cited paper)
Rustrom (2002) [26]	U.S. students (110)	Creative task ('tower of Hanoi')	• Two forms of external enforcement (a penalty or a reward) • Different levels of the external enforcement (none, weak, strong)	Penalties degrade performance; large rewards induce better performance than small (but no better than the no-incentive treatment)	Explicit incentives have a detrimental effect on performance, but only in the case of penalties, not in the case of rewards. Penalties 'distract' subjects.
Gneezy and Rustichini (2000b) [27]	Israeli students (160 for the main experiment)	50 IQ test questions (plus a principal agent game with this framing)	• Different levels of monetary rewards for correct IQ test response (very low, low, high and none)	A discontinuity in the effect of incentives at zero. Small rewards degrade performance; large rewards enhance it.	The presence of the incentive substitutes extrinsic for intrinsic motivation). Categorical crowding out. See Gneezy (2003) [25] **F**
Gneezy and Rustichini (2000b) [28]	Israeli students (180)	Collected donations from households	• Different levels of monetary rewards for the voluntary work (low, high and none)	Discontinuity at zero. Performance with small rewards is lower than performance with high rewards and both are lower than performance with no rewards.	The presence of the incentive substitutes extrinsic for intrinsic motivation). Categorical and Marginal crowding out. See Gneezy (2003) and Upton (1974) [25]

Table 5. Incentives alter the environment in which new preferences are learned (E)

Citation	Subjects (number)	Games or activities	Institutional environments (treatments)	Results relevant to separability	Comment (quotes are from the cited paper)
Burks, Carpenter and Goette (2008) [29]	Swiss and U.S bike messengers (252)	Sequential prisoners' dilemma	• Messenger exposure to performance based pay or not	In a restricted sample unlikely to be affected by selection bias, second movers', exposure to performance pay is associated with between 12 and 15 percent greater likelihood of defection on a cooperative first mover.	The fact that the effects are from a game having no obvious connection with the job suggests that preferences learn under the incentive conditions of the work place are adopted outside the workplace.
Reeson and Tisdell (2008) [30]	Australian Students (98)	Public goods game	• Three external forms of enforcement: - a moral suasion in the form of a single sentence to the effect that the payoff to all would be higher if all contributed (all periods); - a minimum contribution unexpectedly introduced during 4 periods and then removed - none	While the regulation is in place (during the middle stage) contributions are significantly higher than in the initial stage in which only suasion occurrs. After the regulation is removed, contributions are 20 percent lower than in the initial stage. The suasion treatment dramatically increases voluntary contributions compared to a no suasion control.	Extrinsic rewards alter subjects preferences (crowding out other regarding preferences) or beliefs (conveying a different idea of the appropriate behavior in this game.) Categorical amd marginal crowding out. **F, C**
Meier (2007) [31]	Swiss students (11.379)	Contributions to two funds to support financially needy other students.	• Matching donations: For a single semester subjects' contributions are not matched or matched • Matching donations at high or low rates. No matching in subsequent periods	Matching increases contributions when they are in force. Those who experience matching subsequently are substantially less likely to make a contribution to either fund; average contributions show a small, insignificant negative net effect of the incentive.	The negative matching effect is probably not due to the information it conveys on the neediness of the funds (larger effect for the smaller matching rate) or to the subjects' desire to compensate for higher matching induced contributions in the treatment period (subjects whose contribution was unaffected by matching also exhibited a negative effect). **F**
Gaechter, Kessler and Konigstein (2008) * [32]	Swiss students (500)	Gift exchange game	• Three external forms of enforcement: a Trust (pure fixed wage contract), a price deduction (i.e., fine) contract and a bonus incentive contract • Stranger and Partner • Different sequences	Under incentive contracts agents choose a self interested best reply (effort) and there is no voluntary cooperation. If the contract is not incentive compatible under the other contracts there is substantial voluntary cooperation. Experiencing incentive contracts reduces voluntary cooperation even after incentives are withdrawn.	Incentives may have a lasting negative effect on voluntary cooperation. **F**

Table 5 continued…

Citation	Subjects (number)	Games or activities	Institutional environments (treatments)	Results relevant to separability	Comment (quotes are from the cited paper)
Henrich, et al. (2005) [33]	Foragers, herders, others in 15 small-scale societies (1128)	Ultimatum game (plus public goods and dictator games)	• Differences between societies in the level of market integration and the potential payoffs to cooperation	Substantial cross cultural co-variation between the degree of market integration (engagement in market exchange) and both average UG offers and the propensity to reject low offers.	Mutually beneficial interactions in market interactions with strangers may support the evolution of cultures of fair-mindedness towards strangers; "*doux commerce*"? Hirschman (1977). **C**
Irlenbusch and Sliwka (2005) * [34]	German students (84)	Gift-exchange game	• Two internal forms of enforcement: The principal can choose to rely on - a trust (pure fixed wage) contract - compensation contract (i.e., a variable piece rate) • Two different sequences for the contracts	Incentives reduce cooperation (i.e. effort level) and the effect persists after the incentive is removed. Where principals are constrained to offer fixed wages the effort levels of agents are considerably higher than when employers can choose an incentive contract.	Incentives (price rate) alter principals' and agents' perception of the situation: "lead agents to adopt an individual maximization frame ... rather than a cooperative frame," "agents have a stronger concern for the principal's wellbeing in the pure fixed wage setting." (p. 23) **F**
Gneezy and Rustichini (2000a) [35]	Parents from ten day care centers in Haifa, Israel		• An explicit enforcement (i.e. fine) is imposed for lateness in six of these centers.	Tardiness doubles in the six treatment centers and persist even after the fine is removed. No change in the four control centers.	The modest fine signal 'how bad' lateness is and/or is perceived as a price of a service and displace a partially ethical frame by a strategic one: "A fine is a price." **I, F, S**
Bohnet, Frey and Huck (2001) [36]	U.S. students (154)	Contract enforcement game (finitely repeated)	• Different legal institutions (low, medium or high contract enforcement probability) • Low contract enforcement in the last rounds for all sessions.	The probability of enforcement and/or the cost of breach in the early rounds have a non-monotonic effect on contract performance in the later rounds: intermediate levels of contract enforcement decrease trustworthiness, low levels and high levels of legal contract enforcement increase trustworthiness.	"if there is enough time for the crowding dynamics to unfold, environments with low contract enforcement can produce outcomes as efficient as high levels of enforcement." (p.141) "by affecting behavior, institutions affect preferences." (p.142) **F**

Table 5 continued…

Citation	Subjects (number)	Games or activities	Institutional environments (treatments)	Results relevant to separability	Comment (quotes are from the cited paper)
Falkinger, et al. (2000) and personal communication from Gaechter 18 February 2008. [37]	Swiss students (196)	Public goods game	• Incentive compatible (Falkinger (1996)) mechanism and no mechanism; • large and small group size; • Interior and corner Nash equilibria.	Subjects implement the self-interested level of contribution under the mechanism, but contribute substantially more than the self interested level in its absence (until late in the 20 period experiments) (e.g. Figure 5). After experiencing the mechanism subjects contribute 26 percent less when it is withdrawn than those who have experienced it.	By rewarding contributions and penalizing shirkers the mechanism may have relieved subjects' sense of moral responsibility and legitimated the pursuit of self interest. The effects persisted after the withdrawal of the mechanism. **F**, **I**
Carpenter, et al. (2008) [38]	U.S students (172)	Public goods game	• Costly punishment: subjects can punish non-cooperators at a cost to themselves • Different team's residual claim (MPCR -marginal per capita return o n the public good) • Different group size	Shirkers are punished by peers and respond by contributing more, even in the last round unless the frequency of reciprocators is too low or the group is too large. High contributors who are punished subsequently contribute less. (Unpublished results not reported in paper).	Altruistically motivated mutual monitoring, by enhancing shame-induced cooperation, supports high levels of team performance. Synergistic effects of social preferences and incentives. **I**

Table 6. Incentives and social preferences as complements (C)

Citation	Subjects (number)	Games or activities	Institutional environment (treatments)	Results relevant to separability	Comment (quotes are from the cited paper)
Herrmann, Gaechter and Thoni (2008) [39]	16 student pools around the world (1120)	Public goods game (Partner)	• Monetary Costly Punishment	Cooperation is higher in the punishment condition. However, the average payoff with the punishment condition is lower than the average without punishment in many countries. Weak norms of civic cooperation and the weakness of the rule of law in a country are significant predictors of antisocial punishment (punish the high contributors), which reduces the net benefits to the group	Punishment is socially beneficial only if complemented by strong social norms of cooperation with strangers. The quality of the formal law enforcement institutions and informal sanctions are complements, "because antisocial punishment is lower in these societies." P. 1367.
Rodriguez-Sickert, Guzman and Cardenas (2008) [40]	Rural Colombians from 5 communities (128)	Common pool resource game	• Three different forms of external enforcement (A fine regime imposed, a fine proposed to the players and rejected or accepted by them, none) • Different levels of external enforcement (low, and high) for the imposed fine	Under all treatments other than the no fine, groups start at high levels of cooperation. Cooperation remains high only when a fine, be it high or low, is in force. If the players reject the fine, cooperation slowly unravels. Presence of low fines prevented unraveling of cooperation.	When fines are rejected, the implied affirmation of social norms may have temporarily increased cooperation; reciprocal preferences (anger at low contributors) may account for the subsequent erosion of cooperation. Small fines enhance unconditional cooperation by relieving cooperators of the need to retaliate against defectors. **I**, **F**
Galbiati and Vertova (2008a) * [41]	Italian students (216)	Public goods game – one shot (and a lottery game)	• Different levels for the minimum level of contribution rule (zero, low and high) • A symmetric incentive structure (a level of contribution less (more) than the minimum contribution could be subject to a penalty (reward)) with low and medium size	Suggested contributions alone do not induce high levels of cooperation. A high minimum contribution with the presence of an incentive ties up individual average contributions independently of the level of the incentive (low or medium) and increases the expectation about others' contributions.	Incentives not only influence material payoffs but also frame recommended high contributions as obligations. Including both implicit and explicit incentives tie up people's behaviors by activating values and/or coordinating individuals' beliefs, gives salience to minimum contribution rules and make them act as focal points for beliefs about others' contributions. Obligations directly affect average beliefs about others' and preferences for cooperation. Categorical crowding in. **F**

Table 6. Continued…

Citation	Subjects (number)	Games or activities	Institutional environment (treatments)	Results relevant to separability	Comment (quotes are from the cited paper)
Gaechter, Nosenzo and Sefton (2008) * [42]	British Students (84)	Gift-Exchange game with 3-members firms (one employer and two employees)	• Employees move sequentially (Employee 1 has pay comparison information (i.e. information about what coworkers earn) and Employee 2 additionally has effort comparison information (information about how co-workers perform) • Employers could offer high wages to both employees, a high wage to Employee 1 only, a high wage to Employee 2 only, and low wages to both	A homogeneous wage does not affect effort when an employee is matched with a lazy co-worker. Reciprocity is more pronounced when the co-worker is hard-working, as effort is strongly and positively related to own wage and when the employer pays unequal wages to the employees. Exposure to pay comparison information in isolation from effort comparison information does not appear to affect reciprocity toward employers	Unequal wages conditional on worker type may induce high levels of reciprocity based effort; unconditional employer generosity fails to recognize the 'deserving' worker, and is not reciprocated. Incentives and social preferences as complements. Workers respond to employers' recognition of their deservingness, not to employer generosity.
Lopez, et al. (2008a) [43]	240 fisherwomen and fishermen Colombia	Public goods game	• Monetary Costly Punishment. After making the decision individual contributions were publicly posted anonymously and subjects could sanction in private other's contribution decisions • plus an external enforcement (announcement of the socially optimal level of contribution) with monitoring • Different levels of external enforcement (low, and high) for the imposed fine • Two different sequences: monitoring and players' sanctioning and vice versa	Community sanctions combined with external enforcement led to nearly perfect contributions and higher earnings Higher individual contributions with monetary sanctions failed to yield higher earnings. See Masclet, et al. (2003)	Individuals do use the ability to sanction others in their group and increases cooperation. External regulation complements community enforcement efforts. "When community members have better information about the behavior of their neighbors than the external regulator they can fine tune external enforcement efforts" P. 15 Crowding in. See Velez, Stranlund and Murphy (2009)
[44] Serra (2008) *	British students (180)	Bribery game (public official-citizen)	• Three different forms of external enforcement (no monitoring; top-down auditing, and an accountability system which gives citizens the opportunity to report corrupt officials)	Under the combined accountability system, fewer officials engage in corruption. The presence of only top-down auditing did not affect the amount of officers who demanded a bribe but induced corrupt officials to demand a higher bribe than no monitoring.	"non-monetary costs activated by the bottom-up component of the combined system had a significant impact on the public officials decision to engage in bribery." P. 17
[45] Falk, Fehr and Zehnder (2006)	Swiss Students (240)	Labor market game (one employer, three workers)	• With and without a minimum wage. • Two different sequences	The introduction of a legal minimum wage affects workers' fairness preferences leading to a rise in their reservation wages (which persists even after the minimum wage has been removed).	"Minimum wages [may] affect [subjects'] fairness perceptions" (p.1376) creating moral "entitlements". Obligations activate and or enhance social preferences. See Galbiati and Vertova (2008b), Galbiati and Vertova (2008a) [11] [41] **F, E**

Table 6. Continued…

Citation	Subjects (number)	Games or activities	Institutional environment (treatments)	Results relevant to separability	Comment (quotes are from the cited paper)
Falk, Gaechter and Kovacs (1999) [46]	Hungarian students (126, 38)	Gift-exchange game	• Stranger and Partner • Two social approval treatments (face to face, social pressure)	Partner treatment increased effort levels; social pressure has little effect. Wage effort relationship (based on reciprocity) is steeper under partner than under stranger.	Repeated interactions provide powerful incentives while enhancing both intrinsic reciprocity motives and concerns for equitable shares (social pressure adds little).
Barr (2001) * [47]	Zimbabwean villagers (602)	Public goods game	• Two external forms of non monetary punishment - Public announcement: each player announces her level of contribution to everyone present in the session - Subjects could make public verbal statements about each other's decisions: lighthearted criticism or the withholding of praise during informal gatherings	After the introduction of the public announcement and public criticism subjects contribute more.	The fact that non-material punishment raises contributions suggests that it induces shame or other social emotions (the best response for a material payoff maximizer were unaffected). See Gaechter and Fehr (1999) and Mulder, et al. (2006). Subjects may contribute in accordance with their obligations defined with reference to the level of contribution that each member would like all community members to choose. **F**
Gaechter and Falk (2002) [48]	Austrian students (116)	Gift-exchange game	• Stranger and Partner	With repetition, effort levels are higher than one shot interaction and some selfish subjects act strategically as reciprocators and choose the minimal effort level in the last period	Repeated interaction strengthens reciprocity norms and induces 'imitated' reciprocity. "The social norm of reciprocity and the repeated game incentives are complementary." (p.18)
Masclet, et al. (2003) [49]	US (96) and French (44) students (140)	Public goods game	• Two external forms of Punishment with different levels of disapproval (from 0 to 10 points received by a subject from any other agent): Monetary punishment (subjects can reduce the monetary payoff of others after observing their decisions) and non monetary punishment (subjects express disapproval of others' decisions with no effect on others' earnings) • Stranger and Partner • Three stages: In the first and third stages without the punishment. In the second stage, with punishment	Both sanctions show higher and similar levels of contributions. Individuals tend to make higher contributions relative to the preceding period the higher punishment they have received and the lower their contribution was relative to the group average. When the device is removed, having previous monetary sanctions show higher contributions than having non monetary sanctions but the cost of enforcing monetary sanctions causes overall earnings to be similar under both sanction treatments.	Cooperation can be enhanced by non monetary sanctions for reasons that are not strategic and may require repeated interaction. It appears that non monetary punishment, while not affecting the best response of a pay off maximizer, nonethless raised contributions by enhancing the salience of social motives like shame or external peer pressure. Guilt may lead individuals who contribute less than the average to increase their contribution levels more than others. Crowding in. See Lopez, et al. (2008b)

References

Aaron, H. J. 1994. "Distinguished Lecture on Economics in Government: Public Policy, Values, and Consciousness." *Journal of Economic Perspectives*, 8:2, pp. 3-21.

Akerlof, G. A. 1984. *An Economic Theorist's Book of Tales*. Cambridge, UK: Cambridge University Press.

Ariely, D., A. Bracha, and S. Meier. 2009. "Doing Good or Doing Well? Image Motivation and Monetary Incentives in Behaving Prosocially." *American Economic Review. Forthcoming*.

Ariely, D., U. Gneezy, G. Loewenstein, and N. Mazar. 2005. "Large stakes and big mistakes." *Working Papers*. Federal Reserve Bank of Boston

Arrow, K. J. 1971. "Political and Economic Evaluation of Social Effects and Externalities," in *Frontiers of Quantitative Economics*. M. D. Intriligator ed. Amsterdam: North Holland, pp. 3-23.

Arrow, K. J. 1972a. "Gifts and Exchanges." *Philosophy and Public Affairs*, 1:4, pp. 343-62.

Arrow, K. J. 1972b. "Some Mathematical Models of Race Discrimination in the Labor Market," in *Racial Discrimination in Economic Life*. A. M. Pascal ed. Lexington: Lexingotno Books, pp. 187-202.

Bar-Gill, O. and C. Fershtman. 2005. "The limit of public policy: endogenous preferences." *Journal of Public Economic Theory*, 7:5, pp. 841-57.

Barr, A. 2001. "Social dilemmas, shame-based sanctions, and shamelessness: experimental results from rural Zimbabwe." Centre for the Study of African Economies Working Paper WPS/2001.11: Oxford University.

Barr, A., et al. 2009. "Homo Æqualis: A Cross-Society Experimental Analysis of Three Bargaining Games." *Discussion Paper Series*:422 Department of Economics, University of Oxford: Oxford.

Becker, G. S. 1996. *Accounting for Tastes*. Cambridge, MA: Harvard University Press.

Benabou, R. and J. Tirole. 2003. "Intrinsic and extrinsic motivation." *Review of Economic Studies*, 70, pp. 489-520.

Benabou, R. and J. Tirole. 2006. "Incentives and Prosocial Behavior." *American Economic Review*, 96:5, pp. 1652-78.

Bentham, J. 1789. *An Introduction to the Principles of Morals and Legislation*. Oxford: Clarenden Press.

Benz, M. and S. Meier. "Do People Behave in Experiments as in the Field? - Evidence from Donations." *IEW - Working Papers*, Vol. iewwp248. Institute for Empirical Research in Economics - IEW.

Bewley, T. F. 1999. *Why wages don't fall during a recession*. Cambridge: Harvard University Press.

Bisin, A. and T. Verdier. 2001. "The Economics of Cultural Transmission and the Dynamics of Preferences." *Journal of Economc Theory*, 97:2, pp. 298-319.

Bliss, C. J. 1972. "Review of R.M. Titmuss, The Gift Relationship: from human blood to social policy." *Journal of Public Economics*, 1, pp. 162-65.

Bohnet, I. and Y. Baytelman. 2007. "Institution and Trust- Implications for Preferences, Beliefs, and Behavior." *Rationality and Society*, 19:1, pp. 99-135.

Bohnet, I., B. Frey, and S. Huck. 2001. "More Order with Less Law: On Contractual Enforcement, Trust, and Crowding." *American Political Science Review*, 95:1, pp. 131-44.

Borges, G. and B. Irlenbusch. 2007. "Fairness crowded out by law: An experimental study of withdrawal rights." *Journal of Instittional and Theoretical Economics*, 163, pp. 84-101.

Bowles, S. 1989. "Mandeville's Mistake: Markets and the Evolution of Cooperation." *Presented to the September Seminar, London.*

Bowles, S. 1998. "Endogenous Preferences: The Cultural Consequences of Markets and Other Economic Institutions." *Journal of Economic Literature*, 36:1, pp. 75-111.

Bowles, S. 2004. *Microeconomics: Behavior, Institutions, and Evolution*. Princeton: Princeton University Press.

Bowles, S. and H. Gintis. 2006. "Social Emotions," in *The Economy as a Complex Evolving System III: Essays in Honor of Kenneth Arrow*. S. Durlauf and L. Blume eds. Oxford: Oxford University Press.

Bowles, S. and S.-H. Hwang. 2008. "Mechanism design when preferences depend on incentives." *Journal of Public Economics*, forthcoming.

Burks, S., J. Carpenter, and L. Goette. 2008. "Performance Pay and Worker Cooperation: Evidence from an artefactual field experiment." *Journal of Economic Behavior & Organization*, in press.

Camerer, C., L. Babcock, G. Loewenstein, and R. Thaler. 1997. "Labor Supply of New York City Cabdrivers: One Day at a Time." *The Quarterly Journal of Economics*, 112:2, pp. 407-41.

Camerer, C. and E. Fehr. 2004. "Measuring Social Norms and Preferences Using Experimental Games: A Guide for Social Scientists," in *Foundations of Human*

Sociality: Economic Experiments and Ethnographic Evidence from Fifteen Small-Scale Societies. J. Henrich, S. Bowles, R. Boyd, C. Camerer, E. Fehr and H. Gintis eds. Oxford: Oxford University Press.

Cameron, J., K. Banko, and W. D. Pierce. 2001. "Pervasive negative effects of rewards on intrinsic motivation: The myth continues." *Behavior Analyst, Special Issue*, 24:1, pp. 1-44.

Cardenas, J. C. 2004. "Norms from outside and inside: an experimental analysis on the governance of local ecosystems." *Forest Policy and Economics*, 6, pp. 229-41.

Cardenas, J. C., J. K. Stranlund, and C. E. Willis. 2000. "Local Environmental Control and Institutional Crowding-out." *World Development*, 28:10, pp. 1719-33.

Carpenter, J., S. Bowles, H. Gintis, and S.-H. Hwang. 2008. "Strong Reciprocity and Team Production: Theory and Evidence." *Journal of Economic Behavior and Organization*, In press.

Carpenter, J., S. Bowles, H. Gintis, and S.-H. Hwang. 2009. "Strong Reciprocity and Team Production: Theory and Evidence." *Journal of Economic Behavior and Organization*, In press.

Cervellati, M., J. Esteban, and L. Kranich. 2008. "Work Values, Endogenous Sentiments and Redistribution." *Working Paper*. University of Bologna, IAE Barcelona and University of Albany.

Deci, E. L., R. Koestner, and R. M. Ryan. 1999. "A Meta-Analytic Review of Experiments Examining the Effects of Extrinsic Rewards on Intrinsic Motivation." *Psychological Bulletin*, 125:6, pp. 627-68.

Deci, E. L. and R. M. Ryan. 1985. *Intrinsic Motivation and Self-Determination in Human Behavior*. New York and London: Plenum Press.

Dickenson, D. and M.-C. Villeval. 2007. "Does monitoring decrease work effort? The complementarity between agency and crowding-out theories." *Games and Economic Behavior*, doi:10.1016/j.geb.2007.08.004.

Ellingsen, T. and M. Johannesson. 2008. "Pride and prejudice: the human side of incentive theory." *American Economic Review*, 98, pp. 990-1008.

Ellingsen, T., M. Johannesson, S. Munkhammar, and J. Möllerström. 2008. "Why Labels Affect Cooperation." Stockholm School of Economics, Ramböll Management and Harvard University.

Falk, A., E. Fehr, and C. Zehnder. 2006. "Fairness perceptions and reservation wages -- the behavioral effects of minimum wage laws." *Quarterly Journal of Economics*:1347-1381.

Falk, A., S. Gaechter, and J. Kovacs. 1999. "Intrinsic motivation and extrinsic incentives in a repeated game with incomplete contracts." *Journal of Economic Psychology*, 20, pp. 251-64.

Falk, A. and M. Kosfeld. 2006. "The Hidden Costs of Control." *American Economic Review*, 96:5, pp. 1611-30.

Falkinger, J. 1996. "Efficient Private Provision of Public Goods by Rewarding Deviations from Average." *Journal of Public Economics*, 62:3, pp. 413-22.

Falkinger, J., E. Fehr, S. Gaechter, and R. Winter-Ebmer. 2000. "A simple mechanism for the efficient provision of public goods." *American Economic Review*, 90:1, pp. 247-64.

Fehr, E. and A. Falk. 2002. "Psychological Foundations of Incentives." *European Economic Review*, 46:4 - 5, pp. 687-724.

Fehr, E. and S. Gachter. 2000. "Cooperation and Punishment in Public Goods Games." *American Economic Review*, 90:4, pp. 980-94.

Fehr, E. and S. Gaechter. 2002a. "Altruistic Punishment in Humans." *Nature*, 415, pp. 137-40.

Fehr, E. and S. Gaechter. 2002b. "Do Incentive Contracts Crowd Out Voluntary Cooperation?". Institute for Empirical Research in Economics. University of Zurich. Working Paper Series.

Fehr, E., A. Klein, and K. M. Schmidt. 2007. "Fairness and Contract design." *Econometrica*, 75:1, pp. 121-54.

Fehr, E. and J. List. 2004. "The hidden costs and returns of incentives: Trust and trustworthiness among CEOs." *Journal of The European Economic Association*, 2:5, pp. 743-71.

Fehr, E. and B. Rockenbach. 2003. "Detrimental effects of sanctions on human altruism." *Nature*, 422:13 March, pp. 137-40.

Fehr, E. and K. M. Schmidt. 2007. "Adding a Stick to the Carrot? The Interaction of Bonuses and Fines." *American Economic Review*, 97:2, pp. 177-81.

Fershtman, C. and A. Heifetz. 2006. "Read My Lips, Watch for Leaps: Preference Equilibrium and Political Instability." *The Economic Journal*, 116, pp. 246-65.

Fischbacher, U., C. Fong, and E. Fehr. 2005. "Fairness, errors, and the power of competition." *Zurich IEER Working Paper #133*.

Frey, B. and R. Jegen. 2001. "Motivation Crowding Theory: A Survey of Empirical Evidence." *Journal of Economic Surveys*, 15:5, pp. 589 - 611.

Frey, B. and A. Stutzer. 2006. "Environmental Morale and Motivation." *Working Paper*, Vol. 17. Center for Research in Economics, Management and Arts (CREMA). Basel.

Frey, B. S. 1993. "Does Monitoring Increase Work Effort? The Rivalry with Trust and Loyalty." *Economic Inquiry*, 31, pp. 663-70.

Frey, B. S. 1997. "A Constitution for Knaves Crowds Out Civic Virtues." *Economic Journal*, 107:443, pp. 1043-53.

Gaechter, S. and A. Falk. 2002. "Reputation or Reciprocity? Consequences for Labour Relation." *Scandinavian Journal of Economics*, 104:1, pp. 1 - 26.

Gaechter, S. and E. Fehr. 1999. "Collective Action as a Social Exchange." *Journal of Economic Behavior and Organization*, 39:4, pp. 341-69.

Gaechter, S., E. Kessler, and M. Konigstein. 2008. "Performance Incentives and the Dynamics of Voluntary Cooperation." *University of Nottingham, School of Economics*.

Gaechter, S., D. Nosenzo, and M. Sefton. 2008. "The Impact of Social Comparisons on Reciprocity." *IZA, Institute for the Study of Labor*.

Galbiati, R. and P. Vertova. 2008a. "Behavioural Effects of Formal Rules." ECONPUBBLICA, Università Bocconi: Milano.

Galbiati, R. and P. Vertova. 2008b. "Obligations and cooperative behavior in public good games." *Games and Economic Behavior*, in press.

Ginges, J., S. Atran, D. Medin, and K. Shikaki. 2007. "Sacred bounds on rational resolution of violent political conflict." *Proceedings of the National Acadamy of Science*, 104:18, pp. 7357-60.

Gneezy, U. 2003. "The W effect of incentives." University of Chicago Graduate School of Business.

Gneezy, U. and A. Rustichini. 2000a. "A Fine is a Price." *Journal of Legal Studies*, 29:1, pp. 1-17.

Gneezy, U. and A. Rustichini. 2000b. "Pay enough or don't pay at all." *Quarterly Journal of Economics*, 115:2, pp. 791-810.

Greene, J. D., et al. 2001. "An fMRI Investigation of Emotional Engagement in Moral Judgement." *Science*, 293, pp. 2105-08.

Heifetz, A., E. Segev, and E. Talley. 2007. "Market design with endogenous preferences." *Games and Economic Behavior*, 58, pp. 121-53.

Henrich, J., et al. 2005. "'Economic Man' in Cross-Cultural Perspective: Behavioral experiments in 15 small-scale societies." *Behavioral and Brain Sciences*, 28, pp. 795-855.

Herrmann, B., S. Gaechter, and C. Thoni. 2008. "Antisocial Punishment Across Societies." *Science*, 319: 7 March 2008, pp. 1362-67.

Heyman, J. and D. Ariely. 2004. "Effort for Payment: A Tale of Two Markets." *Psychological Science*, 15:11, pp. pp. 787-93.

Hirschman, A. O. 1977. *The passions and the interests : political arguments for capitalism before its triumph*. Princeton, N.J.: Princeton University Press.

Hoffman, E., K. McCabe, K. Shachat, and V. L. Smith. 1994. "Preferences, Property Rights, and Anonymity in Bargaining Games." *Games and Economic Behavior*, 7:3, pp. 346-80.

Hopfensitz, A. and E. Reuben. 2006. "The importance of emotions for the effectiveness of social punishment." *Tinbergen Institute Working Paper 05-0571*

Houser, D., E. Xiao, K. McCabe, and V. Smith. 2007. "When Punishment Fails: Research on Sanctions, Intentions, and Non-Cooperation." *Games and Economic Behavior*, in press.

Irlenbusch, B. and G. K. Ruchala. 2008. "Relative rewards within team-based compensation." *Labour Economics* 15 pp. 141-67.

Irlenbusch, B. and D. Sliwka. 2005. "Incentives, Decision Frames and Motivation Crowding Out- An experimental Investigation." *IZA Discussion paper No 1758*

Laffont, J. J. and M. S. Matoussi. 1995. "Moral Hazard, Financial Constraints, and Share Cropping in El Oulja." *Review of Economic Studies*, 62:3, pp. 381-99.

Levitt, S. D. and J. List. 2007. "What do laboratory experiments measuring social preferences tell us about the real world?" *Journal of Economic Perspectives*, 21:2, pp. pp. 153-74.

Li, J., E. Xiao, D. Houser, and P. R. Montague. 2008. "Neural responses to sanction threats in two-party exchanges." *Baylor College of Medicine*.

Lopez, M. C., J. Murphy, J. Spraggon, and J. K. Stranlund. 2008a. "Does Government Regulation Complement Existing Community Efforts to Support Cooperation? Evidence from Field Experiments in Colombia." *School of Environmental and Rural Studies, Pontificia Universidad Javeriana, Bogotá, Colombia. Department of Resource Economics, University of Massachusetts-Amherst; Department of Economics, University of Alaska-Anchorage*.

Lopez, M. C., J. K. Stranlund, J. Murphy, and J. Spraggon. 2008b. "Comparing the Effectiveness of Regulation and Individual Emotions to Enhance Cooperation: Experimental Evidence from Fishing Communities in Colombią." *School of Environmental and Rural Studies, Pontificia Universidad Javeriana, Bogotá, Colombia. Department of Resource Economics, University of Massachusetts-Amherst; Department of Economics, University of Alaska-Anchorage*.

Lucas, R. E. J. 1976. "Econometric Policy Evaluation: A Critique." *Carnegie-Rochester Conference Series on Public Policy*, Vol. 1, 19-46.

Masclet, D., C. Noussair, S. Tucker, and M.-C. Villeval. 2003. "Monetary and Non-monetary Punishment in the Voluntary Contributions Mechanism." *American Economic Review*, 93:1, pp. 366-80.

Meier, S. 2007. "Do Subsidies Increase Charitable Giving in the Long Run? Matching Donations in a Field Experiment." *Journal of the European Economic Association*, 5:6, pp. 1203-22.

Mellstrom, C. and M. Johannesson. 2008. "Crowding out in blood donation: Was Titmuss right?" *Journal of The European Economic Association*, in press.

Mill, J. S. 1844. *Essays on Some Unsettled Questions of Political Economy, Essay V, Ch. 3* London: John W. Parker.

Mulder, L. B., E. van Dijk, D. De Cremer, and H. A. M. Wilke. 2006. "Undermining trust and cooperation: The paradox of sanctioning systems in social dilemmas." *Journal of Experimental Social Psychology*, 42:147-162.

Reeson, A. F. and J. G. Tisdell. 2008. "Institutions, motivations and public goods: An experimental test of motivational crowding." *Journal of Economic Behavior & Organization.*, 68:1, pp. 273-81.

Rege, M. and K. Telle. 2004. "The impact of social approval and framing on cooperation in public goods situations." *Journal of Public Economics*, 88:7-8, pp. 1625-44.

Rigdon, M. 2009. "Trust and Reciprocity in Incentive Contracting." *Journal of Economic Behavior & Organization* in press.

Rilling, J. K., et al. 2002. "A Neural Basis for Social Cooperation." *Neuron*, 35, pp. 395-405.

Rodriguez-Sickert, C., R. A. Guzman, and J. C. Cardenas. 2008. "Institutions influence preferences: evidence from a common pool resource experiment." *Journal of Economic Behavior & Organization*, in press.

Rosenthal, E. 2008. "Motivated by a Tax, Irish Spurn Plastic Bags." *New York Times*: New York.

Ross, L. and R. E. Nisbett. 1991. *The Person and the Situation: Perspectives of Social Psychology*. Philadelphia: Temple University Press.

Ross, L. and S. M. Samuels. 1993. " The predictive power of personal reputation vs labels and construal in the prisoner's dilemma game." *Stanford university (unpublished) reported in Ross and Ward 1995.*

Ross, L. and A. Ward. 1996. "Naive realism in everyday life: Implications for social conflict and misunderstanding," in *Values and Knowledge*. E. Reed, E. Turiel and B. Terrance eds. Hillsdale, NJ: Lawrence Erlbaum, pp. 103-35.

Rustrom, E. E. 2002. "Sparing the Rod Does not Spoil the Child: An Experimental Study of Incentive Effects." *Moore School of Business, University of SouthCarolina*.

Salant, Y. and A. Rubinstein. 2007. "Choice with Frames." *School of Economics, Tel Aviv University*.

Sanfey, A., et al. 2003. "The Neural Basis of Economic Decision-Making in the Ultimatum Game." *Science*, 300, pp. 1755-58.

Schotter, A., A. Weiss, and I. Zapater. 1996. "Fairness and Survival in Ultimatum and Dictatorship Games." *Journal of Economic Behavior and Organization*, 31:1, pp. 37-56.

Serra, D. 2008. "Combining Top-down and Bottom-up Accountability: Evidence from a Bribery Experiment." Oxford.

Shinada, M. and T. Yamagishi. 2007. "Punishing free riders: Direct and indirect promotion of cooperation." *Evolution and Human Behavior*, 28, pp. 330-39.

Sobel, J. 2007. "Do Markets Make People Selfish?". Department of Economics, University of California: San Diego.

Solow, R. 1971. "Blood and Thunder: Review of The Gift Relationship: From Human Blood to Social Policy by Richard M. Titmuss." *The Yale Law Journal*, 80:8, pp. 1696-711.

Stanca, L., L. Bruni, and L. Corazzini. 2007. "Testing Theories of Reciprocity: Do Motivations Matter?" *Working Papers. University of Milano-Bicocca, Department of Economics.*

Stiglitz, J. 1987. "The Causes and Consequences of the Dependence of Quality on Price." *Journal of Economic Literature*, 25:1, pp. 1-48.

Taylor, M. 1987. *The possibility of cooperation*. New York: Cambridge University Press.

Tenbrunsel, A. and D. M. Messick. 1999. "Sanctioning systems, decision frames and cooperation." *Administrative Science Quarterly*, 44, pp. 684-707.

Tirole, J. 1999. "Incomplete Contracts: Where do we stand?" *Econometrica*, 67:4, pp. 741-78.

Tversky, A. and D. Kahneman. 1981. "The framing of decisions and the psychology of choice." *Science*, 211:4481, pp. 453-58.

Tyran, J.-R. and L. Feld. 2006. "Achieving Compliance when Legal Sanctions are Non-deterrent." *Scandinavian Journal of Economics*, 108:1, pp. 135-56.

Upton, W. E. I. 1974. "Altruism, attribution, and intrinsic motivation in the recruitment of blood donors." *Dissertation Abstracts International*, 34:12, pp. 6260-B.

Velez, M. A., J. K. Stranlund, and J. J. Murphy. 2009. "Centralized and Decentralized Management of Local Common Pool Resources in the Developing World: Experimental Evidence from Fishing Communities in Colombia." *Economic Inquiry*, forthcoming.

Young, P. and M. Burke. 2001. "Competition and Custom in Economic Contracts: A Case Study of Illinois Agriculture." *American Economic Review*, 91:3, pp. 559-73.

Zajonc, R. B. 1968. "Attitudinal Effects of Mere Exposure." *Journal of Personality and Social Psychology Monograph Supplement*, 9:2, Part 2, pp. 1-27.

Zhong, C.-B., J. Loewenstein, and J. Murnighan. 2007. "Speaking the same language: the cooperative effects of labeling in the Prisoners' Dilemma." *Journal of Conflict Resolution*, 51, pp. 431- 56.