UNIVERSITÀ DEGLI STUDI DI SIENA

QUADERNI DEL DIPARTIMENTO
DI ECONOMIA POLITICA E STATISTICA

Samuel Bowles
Sandra Polania-Reyes

Economic incentives and social preferences: substitutes or complements?

n. 617 – Ottobre 2011

Abstract - Explicit economic incentives designed to increase contributions to public goods and to promote other pro-social behavior sometimes are counterproductive or less effective than would be predicted among entirely self-interested individuals. This may occur when incentives adversely affect individuals' altruism, ethical norms, intrinsic motives to serve the public, and other social preferences. In the 50 experimental studies that we survey these effects are common, so that incentives and social preferences may be either substitutes (crowding out) or complements. We provide evidence for four mechanisms that may account for these incentive effects on preferences, based on the fact that incentives may (i) provide information about the person who implemented the incentive, (ii) frame the decision situation so as to suggest appropriate behavior, (iii) compromise a control averse individual's sense of autonomy and (iv) affect the process by which people learn new preferences. An implication of the fact that incentives affect preferences is that the evaluation of public policy must be restricted to allocations that are supportable as Nash equilibria when account is taken of these crowding effects. We show that well designed fines, subsidies and the like minimize crowding out and may even do the opposite, making incentives and social preferences complements rather than substitutes.

JEL codes: A13 (Relation of Economics to Social Values); C90 (Experiments); D02 (Institutions); D63 (Equity, Justice, Inequality, and Other Normative Criteria and Measurement); D64 (Altruism); H41 (Public goods); D78 (Policy making and implementation); E61 (Policy Objectives; Policy Designs and Consistency); Z13 (Social norms and social capital)

Keywords: Public goods, behavioral experiments, social preferences, endogenous preferences, motivational crowding, explicit incentives

For their contributions to this paper we thank Mahzarin Banaji, Abigail Barr, Iris Bohnet, Stephen Burks, Juan Camilo Cardenas, Jeff Carpenter, David Echeverry, Ernst Fehr, Urs Fischbacher, Christina Fong, Simon Gächter, Roberto Galbiati, Joshua Greene, Sung-Ha Hwang, Jonathan Haidt, Daniel Houser, Steffen Huck, Bernd Irlenbusch, Magnus Johannesson, Esther Kessler, Nicola Lacetera, Maria Claudia Lopez, Thomas Schelling, Carlos Sickert-Rodriguez, David Skuse, Joel Sobel, John Stranlund, Elisabeth Wood as well as three anonymous referees and the editor. Financial support by the Behavioral Sciences Program of the Santa Fe Institute, the U.S. National Science Foundation, the University of Siena and the European Science Foundation is gratefully acknowledged.

Samuel Bowles, Santa Fe' Institute and University of Siena
Sandra Polania-Reyes, University of Siena and University College London

1. Introduction

Thomas Schelling recalls his "exciting and stimulating times" in the early 1950s as a young staffer 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 at the White House 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…? Saturday meetings virtually disappeared."

Was Schelling's experience atypical? Incentives work, often affecting the targeted behavior almost exactly as conventional economic theory predicts: textbook examples include the work response of Tunisian sharecroppers and American windshield installers, and the reduced criminal activities of former Italian convicts who could expect more severe sentences if convicted (Laffont and Matoussi (1995), Lazear (2000), Drago, Galbiati, and Vertova (2009).

But explicit economic incentives sometimes have surprisingly limited effects and may even be counterproductive. Substantial rewards for high school matriculation in a randomized experiment in Israel had no impact on boys and little effect on girls except among those already quite likely to matriculate (Angrist and Lavy (2009)). Large and in most cases immediate cash payment in return for tested scholastic achievement in 250 urban schools in the U.S. were almost entirely ineffective, while incentives for student inputs (reading a book, for example) had the intended, if modest effects (Fryer (2011)). In an unusual natural experiment, the imposition of fines designed to reduce hospital stays in Norway had the opposite effect (Holmås, Kjerstad, Lurås, et al. (2010)) while in England hospital stays were greatly reduced by a policy designed to evoke shame and pride in hospital managers rather than the calculus of profit and loss (Besley, Bevan, and Burchardi (2009)).

Anecdotal accounts of what appear to be even more dramatic cases of counterproductive incentives are common. On December 1, 2001 the Boston Fire Department terminated its policy of unlimited paid sick days, replacing it with a 15-day sick day limit; pay would be docked for firemen exceeding the limit. The firemen responded to the new incentives: those calling in sick on Christmas and New Year's Day increased tenfold over the previous year. The Fire Commissioner retaliated by cancelling their holiday bonus checks (Belkin (2002)). The firemen were unimpressed: the year following they claimed 13,431 sick days; up from

6,432 the previous year (Greenberger (2003)). Many of the firemen, apparently angered by the new system, abused it or abandoned their previous ethic of serving the public even when injured or not feeling well.

Not surprisingly, then, since Richard Titmuss' *The Gift Relationship: From Blood Donations to Social Policy,* economists have been intrigued by the claim that policies based on explicit economic incentives may be counterproductive when they induce people to adopt what Titmuss called a 'market mentality' or in some other way compromise pre-existing values to act in socially beneficial ways. But few were persuaded (Solow (1971), Arrow (1972), Bliss (1972)).

At the time of its publication there were two strong reasons to doubt Titmuss' claim: there was little hard evidence that social preferences are important influences on individual behavior; and there was even less evidence (in the Titmuss (1971) book or elsewhere) that social preferences would be undermined by explicit economic incentives (which we will call simply "incentives" without the adjectives, meaning interventions to influence behavior by altering the economic costs or benefits of some targeted activity.)

Theoretical and empirical advances over the intervening years provide the basis for a reconsideration of these issues (Kreps (1997), Elster (1998), Rabin (1998), Loewenstein (2000), Sobel (2002)). First, evidence from both the behavioral experimental laboratory and the field is consistent with the view that social preferences are important influences on economic behavior (Fehr, Gächter, and Kirchsteiger (1997), Bewley (1999), Fehr and Schmidt (1999), Fehr and Gächter (2000), Young and Burke (2001), Bandiera, Barankay, and Rasul (2005), Falk, Fehr, and Fischbacher (2005), Fehr, Klein, and Schmidt (2007), DellaVigna (2009), Leider, Möbius, Rosenblat, et al. (2009), Sloof and Sonnemans (2011)).

Second, the importance of incomplete contracts has been widely recognized in theoretical works 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 referring to the social norms and identities that underwrite mutually beneficial exchange in the absence of complete contracts (Arrow (1971), Becker (1976), Akerlof (1984), Helsley and Strange (2000), Benabou and Tirole (2006), MacLeod (2007), Sliwka (2007), Ellingsen and Johannesson (2008), Akerlof and Kranton (2010)).

Finally, advances in the theory of public policy have addressed cases in which incentives affect both beliefs and preferences and may thus have unintended effects (Lucas

(1976), Taylor (1987), Bowles (1989), Aaron (1994), Sunstein (1996), Frey (1997), Cooter (2000), Bowles (2004), Bar-Gill and Fershtman (2005), Sobel (2005), Cervellati, Esteban, and Kranich (2010))

2. Overview: Incentives and social preferences as substitutes or complements

We use the term "social preferences" to refer to motives such as altruism, reciprocity, intrinsic pleasure in helping others, inequity aversion, ethical commitments and other motives that induce people to help others more than would an own-material-payoff maximizing individual. Our use of the term is thus not restricted to cases in which the actor assigns some value to the payoffs received by another person, as in the utility functions of Fehr and Schmidt (1999), Rabin (1993) and Levine (1998). While these functions provide a convenient way to model some of the motivations for pro-social behavior, we use the broader definition because moral, intrinsic, or other reasons unrelated to a concern for another's payoffs often motivate people to help others, adhere to social norms, and act in other pro-social ways even when it is personally costly to do so. A person, for example, may adhere to a social norm not because of the harm that a transgression would do to others, but because of the kind of person she would like to be; helping the homeless may be motivated by Andreoni's "warm glow" of giving rather than a concern with the wellbeing of the poor (Andreoni (1990)).

The standard (if generally implicit) assumption in economics is that the behavioral functions relevant for mechanism design, public economics and related fields are separable in social preferences (should they exist) and incentives. This means, for example, that the citizen's response to variations in a subsidy for contributions to a public good is independent of her pre-existing level of social preferences. It also means that the effect of variations in her pre-existing non-economic motivations on the citizen's level of contributions does not depend on the presence or magnitude of incentives.

We call this the separability assumption. It implies that taxes, subsidies, and other incentives affect behavior only by altering the economic costs and benefits of the targeted activities. But where the separability assumption does not hold, social preferences may be either heightened by incentives appealing to self-interest or, the more commonly observed case, affected adversely by incentives.

This is illustrated in Figure 1 where, due to the effect of incentives on preferences, the total – direct and indirect -- effect of the incentive may fall short of that which works directly on the costs and benefits of the targeted activity. In this case we say that incentives crowd out social preferences and that incentives and social preferences are substitutes: the effect of each

on the targeted activity declines, the greater is the level of the other. Where the effect on social preferences is positive, crowding in occurs and social preferences and incentives are complements, the level of each enhancing the effect of the other.

[Figure 1 here]

The possibility that incentives designed for material payoff-maximizers might have adverse effects is a familiar theme in political science (Taylor (1987), Grant (2011)), psychology (Deci (1975)), sociology (Healy (2006)), and the other social sciences; but it has found few adherents in economics. The reason is that we have adopted a simplifying strategy that goes back at least to John Stuart Mill (1867[1848]): 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.

In other words, we ignore the two lower arrows in Figure 1. But recent experimental and other evidence has prompted many economists to reconsider Mill's simplification. To further this reconsideration we here provide a taxonomy of incentive effects on preferences based on two distinctions: their nature and their causes.

Concerning the first, people often react to the mere presence of incentives rather than their extent (Gneezy (2003)): giving to charity when tax breaks are involved (whatever their magnitude) may feel different or send a different signal than would be the case in the absence of these incentives. But the extent of the incentive may also matter. Thus the effects of incentives on social preferences may be either marginal (depending continuously on the level of the incentive) or categorical (the presence of incentives affecting social preferences independently of their level) or a combination of the two.

We also distinguish between two causes of incentive effects on preferences. First, behavior is acutely sensitive to the nature of the decision situation (Ross and Nisbett (1991), Tversky and Kahneman (1981)); and, as we will see, the presence or extent of incentives provides information about the situation. A psychologist might say that preferences are "situation-dependent" and that incentives provide situational clues. We say that the preferences are state-dependent, with differing incentives constituting different states. In the next section we offer a model of incentive-state-dependent preferences and provide data indicating that both categorical and marginal crowding out occurs.

State dependence arises because actions are motivated by a heterogeneous repertoire of preferences –from spiteful to payoff-maximizing to generous, for example -- the salience of which depends on the nature of the decision situation – interacting with a domineering supervisor, shopping, or relating to one's neighbors, for example. To see how this works, think about gifts. Economists know that money is the perfect gift – it replaces the giver's less well-informed choice of a present by the recipient's own choice. But when the holidays come around few economists give money to their friends, family and colleagues. This is because we also know that money cannot convey thoughtfulness, concern, whimsy, or any of the other messages that non-monetary gifts sometimes express. A gift, we know, is more than a transfer of resources; it is a signal about the giver and her relationship to the recipient, and money changes the signal.

Can the same be said of incentives? A long tradition in psychology has concluded that it can:

> The multiple meanings of ... tangible rewards are reflected in our everyday distinction among bribes and bonuses, incentives and salaries. … they carry different connotations concerning, for example, *(i)* the likely conditions under which the reward was offered, *(ii)* the presumed motives of the person administering the reward, and *(iii)* the relationship between the agent and the recipient of the reward (Lepper, Sagotsky, Dafoe, et al. (1982) numbers added).

As Mark Lepper and his coauthors say, incentives may affect preferences for a reason familiar to economists, that is because they indicate "the presumed motives of the person administering the reward." By implementing an incentive, a principal reveals information about his or her intentions (own payoff maximizing vs. fair-minded, for example) as well as beliefs about the target of the incentives (hardworking or not) and the targeted behavior (how onerous it is, for example.) This information, in turn, may then affect the target's non-economic motivation to undertake the task at hand. In section 5 we present experimental evidence that the information provided about the principal can sometimes attenuate or even reverse the intended effect of the incentive. Of course when an incentive provides good news about the principal's intentions or type—when rewards are offered, for example, rather than fines– it may recruit the target's social preferences to work synergistically with the direct effect of the incentive on the net material benefits to the agent of taking the targeted action. In this case incentives and social preferences become complements rather than substitutes. We will see (in section 9 and 11) that this crowding in phenomenon is sometimes observed in

experiments, for example, when the principals implementing incentives are peers in a public goods game who pay to fine free riders in order to support cooperative norms.

But there are other reasons, less familiar to economists, for state-dependence: reasons that do not concern information about the principal, and that may be at work even in non-strategic settings. A second mechanism is that incentives provide cues about (as Lepper and his coauthors put it) "the likely conditions under which the reward was offered:" by framing a decision situation, economic incentives may provide cues for appropriate behavior. This second mechanism is distinguished from the first in the experimental evidence by the fact that in the former the incentives are implemented by a principal who is a player in the game; while in the latter the targets of the incentive are not playing against the incentive designer; rather the incentives are introduced by the experimenter.

Situational cues may be very subtle, and our responses to them unwitting. When experimental subjects had the opportunity to cheat on a test and as a result to gain higher monetary rewards, less than quarter did so when the room was brightly lit, but more than half cheated when the room was slightly less well lit (the variations in lighting had no effect on the observability of cheating.) In another experiment subjects who wore (nonprescription) dark glasses were much less generous to their partner in a Dictator Game than were those outfitted with clear glasses (Zhong, Gino, and Bohns (2010)). The dark glasses and darkened room gave the subjects a sense of anonymity, the researchers found. But it was entirely illusory: it is difficult to imagine that a subject could really think that his own wearing dark glasses would make him less observable, especially given that the experiment was conducted at computer terminals in closed cubicles.

The degree of anonymity differs dramatically as we move between family, workplace, marketplace and other domains of social interaction. Fiske (1992) provides a taxonomy of four psychological models corresponding to distinct kinds of social relationships: authoritarian, communal, egalitarian and market, each with culturally prescribed patterns of appropriate behavior. Depending on the information they convey, incentives may signal that the situation corresponds to any one of these four types, and therefore evoke distinctive responses.

We will see that a plausible explanation of some of the framing effects of incentives observed in experiments is that it occurs because market-like incentives trigger what psychologists term "moral disengagement" (Bandura (1991), a process that occurs because "people can switch their ethicality on and off" (Shu, Gino, and Bazerman (2009):31). In section 6 we review experiments in which crowding out appears to have been the result of

moral disengagement. Depending on the information they convey, incentives may also trigger the opposite – moral engagement – and, as we will see in section 9, experiments provide a few examples of this form of crowding in, illustrating the possible synergy or complementarity between social preferences and incentives.

The third mechanism that makes social preferences state dependent is the crowding out of intrinsic motives by incentives (or constraints) that compromise a subject's sense of autonomy (Deci and Ryan (1985), Deci, Koestner, and Ryan (1999)). These effects may occur in strategic situations where the bad news that incentives convey concerns the desire of a principal to control the agent. But most of the experimental evidence for this third crowding out mechanism comes from non-strategic settings (the experimenters, not a principal implements the incentive.) The underlying psychological mechanism appears to be a fundamental desire for "feelings of competence and self-determination" that are associated with intrinsically motivated behavior (Deci (1975)).

According to this interpretation, where people derive pleasure from an action *per se* in the absence of other rewards, the introduction of incentives may 'over-justify' the activity and reduce the individual's sense of autonomy. This self-determination mechanism differs from the previous two mechanisms -- bad news about a principal and moral disengagement– because it arises from the target's desire for autonomy and does not depend on the target inferring negative information about a principal or clues about appropriate behavior. This is particularly evident in some early "over-justification" experiments in which when a financial reward was offered by the experimenter, children often forsook previously uncompensated activities in which they had enthusiastically engaged, like painting. More recent experiments show the same negative effects of incentives on altruistic behavior (Warneken and Tomasello (2008)). In the absence of rewards kids less than two years old avidly helped an adult retrieve an out of reach object; but after being rewarded with a toy for their helping behavior the helping rate fell by forty percent.

The fact that the incentive was a reward rather than a penalty suggests that it did not convey negative information about the incentive designer, but instead altered the meaning of the activity itself from one that expressed autonomy to one that expressed compliance. The interpretation that self-determination is involved in the negative response to incentives is consistent with the fact that close supervision or arbitrary temporal deadlines for completion of an otherwise enjoyable activity have effects very similar to financial rewards (Lepper, et al. (1982)). In section 7 we survey experimental evidence for this "control aversion" mechanism for state dependent preferences.

We have just described three (partially overlapping) reasons why the state dependent nature of preferences might lead to crowding out. For ease of reference we will call them "bad news," "moral disengagement," and "control aversion." But in addition to incentives altering the preferences that motivate an individual's action by altering the subject's sense of the situation, there is a second and quite different way that incentives may affect preferences. The type and extent of a society's use of economic incentives also may affect the process of preference-updating by which individuals acquire new tastes or social norms that will persist over long periods. Models from biology, anthropology, and economics allow us to formalize this learning process (Cavalli-Sforza and Feldman (1981), Boyd and Richerson (1985), Guth and Kliemt (1994), Bowles (1998), Bisin and Verdier (2011), Bowles and Gintis (2011)).

The key difference between endogenous and state-dependent preferences is that in the former case the effect of the incentive on preferences persists in the long run because the updating process on which cultural transmission is based typically occurs during youth and its effect endures over decades if not entire lifetimes. We say that incentives affect preferences in both the state-dependent and endogenous preference case, but the mechanism of the effect is different: in the former case the incentive is a reversible signal about the principal or the situation, in the latter the incentive alters the preference-updating process.

An example unrelated to incentives may clarify the difference between endogenous and state-dependent preferences. As Italian residents, your authors now eat a lot more pasta than we did in our countries of origin. Abstracting from possible international price differences, this could be another case of "when in Rome, do as the Romans." Or it might be that we have newly come to enjoy the taste of pasta, perhaps through extensive exposure to it while in Italy. Which case it is – state-dependent or endogenous preferences – would be revealed by what we will eat back in Bogotá or Santa Fe. If we go back to *arepas* or potatoes, then our taste for pasta was state-dependent. If we remain *pastaphiles*, then our preferences have endogenously changed.

Preferences may be endogenous in this sense because the extent to which a society relies on economic incentives – as opposed to other kinds of motivations and controls – may affect how people learn new preferences (evidence for the endogeneity of preferences is surveyed in Bowles (1998) and (2004), Bowles and Gintis (2011).) The learning on which preference endogeneity is based is of course a long term process unlikely to be observed in a brief experiment. Nonetheless experiments may provide clues that learning is affected: we take as evidence consistent with preference endogeneity those cases in which crowding

effects of incentives persist after the removal of the incentive. (Other explanations not involving endogenous preferences are generally also possible in these situations.)

In sections 3 and 4 we make explicit the underlying causal mechanisms through the use of models of state-dependent and endogenous preference formation, Table 1 summarizes the differences.

[Table 1 here]

Our empirical strategy (based on experimental results) is to observe the total effect of incentives on behavior and to note whether this differs from the predicted direct effect (the top arrows in Figure 1) in order to infer the effects of incentives on (unobserved) social preferences and thereby on actions (the bottom two arrows). Our data set includes all the economic experiments we have been able to locate that allow a test of the separability assumption. Our tables include more than a hundred different subject pools, over twenty-six thousand subjects from 36 countries, playing Dictator, Trust, Ultimatum, Public Goods, Third Party Punishment, Common Pool Resource, Gift Exchange and other principal-agent games. These are all settings in which one's actions affect the payoffs of others so that social preferences may affect a subject's experimental behavior. We find evidence of non separability in all of these games. Because non separability, as we will see, arises from the social relationships among those imposing incentives and their targets and the nature of the incentive, and because game structures differ in this respect, it would be surprising if the nature and degree of non separability did not differ across these games. However, lacking a metric for non-separability that is comparable across games, we have not explored this possibility.

Few experiments have thus far been designed to address the causes of non-separability, so the inferences that we draw 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. We refer to a number of important experiments done using other methods in the text. All reported results are statistically significant at conventional levels unless noted.

Incentives may have counter-intuitive and counterproductive effects for reasons other than non-separability (Seabright (2009)). Strong monetary incentives, for example, may over-motivate an agent leading to greater than the optimal level of arousal. This appears to be the

mechanism underlying the negative effects of high incentives found in three experiments by Ariely, Gneezy, Loewenstein, et al. (2009). In other cases incentives alter individuals' beliefs about the actions of others, with possibly counter-intended effects. This is often thought to be the case when tax authorities announce stiff penalties for underpayment, unwittingly letting the public know that that cheating is common and thereby promoting rather than deterring it. We do not consider these and other cases of counterproductive incentives where the mechanisms are unrelated to the non-separability of incentives and social preferences, which is the focus of this paper.

Some of the experimental results presented below may be explained by more than one of our four mechanisms that account for non-separability, either because the mechanisms are not mutually exclusive so that multiple mechanisms are at work, or because the experiment does not provide sufficient information to say which mechanism accounts for the evidence of non-separability. In these ambiguous cases we classified the experiment as an illustration of the mechanism which we thought best accounted for for the crowding result that we report (We indicate in each table where such ambiguities occur and which other mechanisms may have been at work). As a preview, Figure 2 presents a summary of our findings, the size of the ellipses indicating the total number of studies that exhibit each of the four crowding out mechanisms in question, and the intersections giving the cases where multiple mechanisms may be involved.

There are two reasons why despite the considerable number of experiments in which preference effects of incentives appear to be at work it is difficult to estimate how prevalent these effects are in real economies. First, the experimental games involved are about social dilemmas or sharing with others, that is, settings in which social preferences are likely to be important and therefore there is something to be crowded out or in. While the experimental evidence suggests that crowding out may affect blood donations or participation in community service projects, it does not have much to say about the effect of incentives on shopping behavior or cleaning hotel rooms. Second, while section 10 presents evidence that experimental play in these social dilemmas predicts behavior in some non-experimental situations, isolating social preferences from other influences on behavior in natural settings is difficult. We conclude in sections 11 and 12 with policy implications.

[Figure 2 here]

3. Incentives as signals: a model of state-dependent preferences

In this section (following Hwang and Bowles (2011a) and Bowles and Hwang (2008)) we model incentive effects on state-dependent preferences and clarify the distinction between categorical and marginal incentive effects by means of an empirical illustration. We consider an individual who may bear a cost to take an action that confers benefits on others, which may be encouraged by a subsidy implemented by a social planner. Citizens also have values that may motivate such pro-social actions even in the absence of the subsidy. We study a single member of a community of identical citizens who may contribute to a public project by taking an action a at a cost $g(a)$ that is increasing and convex in its argument, and that may be offset partially by a subsidy s, that is proportional to the individual's level of contribution. The output of the project is available in equal measure to all, and it varies positively and linearly with A, the sum of the n members' contributions, according to φA where φ is a positive constant.

We express the individual's social preferences as v, the effect of an increase in the contribution level on the individual's utility that is unrelated to material payoffs. Thus we have the individual's utility

(1) $u = \varphi A - g(a) + as + av$

we make explicit the sources of non-separability by the value function:

(2) $v(s; \lambda_0, \lambda_c, \lambda_m) = \lambda_0(1 + \mathbf{1}\{s > 0\}\lambda_c + s\lambda_m)$

where the indicator $\mathbf{1}\{s > 0\} = 1$ if $s > 0$ and zero otherwise. In equation (2) $\lambda_0 \geq 0$ measures the citizen's baseline social preferences namely the citizens values in the absence of a subsidy or $v(s; \mathbf{0})$, λ_c (which may be of either sign) measures the categorical effect of the presence of an incentive, and λ_m (which also may be of either sign) measures the marginal effect of variations in s on values for $s > 0$. The crowding effects represented by λ_c and λ_m in (2) may arise because of any of the three mechanisms by which state dependent preferences arise: bad news, moral disengagement or control aversion.

The individual's utility is thus

(3) $u = \varphi A - g(a) + a\big(s + \lambda_0(1 + \mathbf{1}\{s > 0\}\lambda_c + s\lambda_m)\big)$

and the individual's utility maximizing contribution (a^*) equates the marginal cost of contributing to the marginal benefits (the returns from the public good plus the subsidy plus the effect on the individual's values), or:

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

We assume that in the absence of a subsidy the contributions of the citizens to the public good given by (4) are inefficient in the sense that there exists a mutual increase in contributions that would make all citizens better off. The causal structure of the model is illustrated in Figure 3.

[Figure 3 here]

The introduction of a subsidy increases contributions by raising the marginal benefits of contributing, the right hand side of (4) which we denote, θ. Considering the case in which there initially is no incentive, the effect of an incentive on the net benefits of contributing (expressed in discrete terms so as to be able to account for the discontinuity in the value function at $s = 0$) is

(5) $$\Delta\theta/\Delta s = 1 + \lambda_0\left(\frac{\lambda_c}{\Delta s} + \lambda_m\right)$$

and is composed (as expected) of a direct effect (that is, 1, the top arrow in Figure 1),and the indirect state-dependent effect which will be negative in the case of crowding out (λ_c or λ_m negative), and larger in absolute value the greater are the baseline values of the individual (λ_0). We likewise see that

(6) $$\Delta\theta/\Delta\lambda_0 = 1 + \mathbf{1}\{s > 0\}\lambda_c + s\lambda_m$$

which, in the case of crowding out, is declining in s.

Equations (5) and (6) make it clear that when λ_c and λ_m are negative, incentives and baseline values are substitutes: the effect of each on the marginal benefits of contributing varies inversely with the level of the other. The fact (from equation 5) that the crowding effect is larger for those with greater baseline social preferences makes sense and is consistent with experiments that have identified the strength of individuals' social preferences independently of incentives and found that crowding out effects are larger for those with greater baseline values (Bohnet and Baytelman (2007), Kessler (2008), Carpenter and Myers (2010)). This substitutability between incentives and baseline values will be important when we address questions of public policy in the penultimate section.

Using (5) we say that a particular change in incentives Δs has crowded out social preferences if $\Delta\theta/\Delta s < 1$, that is, if the total effect of the incentive is less than the direct effect, and conversely for the case of crowding in. Crowding will not occur if λ_c and λ_m or λ_0 are zero (that is, if social preferences are not state dependent, or they are absent). What we term strong crowding out holds if $\Delta\theta/\Delta s < 0$, which can occur if categorical crowding out is large relative to the size and marginal effect of the subsidy, or if the marginal effect is

negative.

The two forms of non-separability are illustrated by crowding out in Figure 4. Crowding in, which we do not show, would either shift the "separability" function upwards – categorical crowding in -- or increase its slope – marginal crowding in. Because the functions in Figure 4 represent the citizens' best responses to the planner's choice of an incentive and thus constitute one of the constraints making up the planner's optimizing problem, we call these functions the planner's implementation technology.

[Figure 4 here]

An experiment allows an estimate of both categorical and marginal crowding out. Bernd Irlenbusch and Gabriele Ruchala 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 (Irlenbusch and Ruchala (2008) details are in Table 2, results are shown in Figure 5). Payoffs were such that even with no incentive individuals would maximize their payoffs by contributing 25 units. In the no-incentive case contributions averaged 37 units, or 48 percent above what would have occurred if the participants had been motivated only by the material rewards of the game. 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 (53 units) barely (and insignificantly) exceeded that predicted for self-interested subjects (50 units).

[Figure 5 here]

In Figure 5 we use the observed behavior in the high and low bonus case along with the assumption that marginal crowding affects the slope of the citizens' best response function by a given amount (so that the function remains linear as in Figure 4) to estimate the marginal effect of the bonus. We find 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, that is, had subjects without social preferences simply best responded to the incentive. Crowding out thus affected a 26 percent reduction in the marginal effect of the incentive. The estimated response to the incentive also gives us the level of categorical crowding out, namely the difference between the observed contributions (37.04) in the

absence of any incentive and the predicted contributions had an arbitrarily small incentive been in effect (the vertical intercept of the observed line in Figure 5) or 34.55. 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, 12.04).

Categorical crowding out is also evident in other experiments. In one, 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 (Heyman and Ariely (2004)). 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).

Another experiment that allows us to distinguish categorical and marginal crowding was implemented by Juan Camilo Cardenas (2004), but here (as in some other experiments) we observe categorical crowding in. Cardenas implemented an experimental Common Pool Resource Game very similar in structure to the kind of real world commons problem faced by his subjects – rural Colombian eco-system users. In the absence of any explicit incentives, the villagers on average extracted 44 percent less of the experimental "resource" than would have maximized their individual payoffs, providing evidence of a significant willingness to sacrifice individual gain so as to protect the resource and raise group-average payoffs. When they were liable to pay a small fine (imposed by the experimenter) if they over-extracted the resource, as expected, they extracted even less than without the fine, showing that the fine had the intended effect.

The fact that the average extraction under the small fine treatment was 55 percent less than the Nash equilibrium for self-interested subjects (when account is taken of the fine) suggests that the fine had increased the salience of the villagers' social preferences (by 25 percent, if the 44% deviation from the self-interested Nash behavior is taken as the measure of social preferences). Interestingly, raising the fine from a low to a high level had virtually no effect. Variations in the fine thus did not work as an incentive, but rather (in Cardenas' view) the very presence of the fine (high or low) was a signal, one that alerted subjects to the public good nature of the interaction. We will present other examples of fines as signals (section 3) and crowding in (section 9). These cases hold important lessons for why incentives sometimes are counterproductive and how well-designed policies can make incentives and social preferences complements rather than substitutes.

Unfortunately, unlike the Irlenbusch and Ruchala and Cardenas studies, many experiments do not establish the response to incentives that would be observed under

separability, so it is impossible to determine if incentives are "under-performing." A common misinterpretation of experimental results is to infer from the observation that an incentive has an effect in the intended direction that crowding out has not occurred (Rigdon (2009)). But observing a positive incentive effect in an experiment does not preclude crowding out. It is clear from Figure 4 (or equation 5 and the definition of crowding out) that a positive incentive effect may occur in the presence of marginal crowding out (as long as it is not "strong") and in the presence of categorical crowding out (as long as the incentive is sufficiently large.) For example, consider some substantial incentive indicated by s^+ in the figure. Under both marginal and categorical crowding out, the action taken (points **a** and **b** respectively) is greater than in the absence of the incentive (**d**), so the incentive "worked": it affected the action in the intended direction. But the diagnostic for the presence of crowding is a comparison of these two action levels with the level that would have occurred under separability, namely point **c**, and this comparison makes it clear that crowding out occurred.

4. Incentives alter cultural learning: a model of endogenous preferences.

A quite different mechanism by which crowding might occur has also been studied, one in which preferences are endogenous so that one or more of the parameters of the individual's value function -- λ_0, λ_c and λ_m-- are altered by incentives (Bar-Gill and Fershtman (2005), Hwang and Bowles (2011b)). Hwang and Bowles present a model of cultural evolution in which the presence or level of incentives affects the process by which preferences are acquired or abandoned, so that a population's equilibrium distribution of preferences depends on incentives. By equilibrium preferences they mean a configuration of incentives and preferences such that the latter are stationary given the process of preference-updating.

In the Sung-Ha Hwang and Bowles model preferences are endogenous because i) schools, families, religious organizations and other societal institutions seek to promote civic minded values and ii) individuals periodically alter their preferences in response to their own recent experiences. Their model of endogenous preferences is based on two empirical regularities. The first is the powerful effect of mere exposure on preferences, documented by the social psychologist Zajonc (1968) and in subsequent works (Birch and Marlin (1982), Murphy and Zajonc (1993), Murphy, Monahan, and Zajonc (1995)). The exposure effect is one of the reasons that cultural transmission may favor the numerous over the rare, independently of their economic success (See Boyd and Richerson (1985):223ff, Ross and Nisbett (1991):30ff, Bowles (1998) and the works cited there.) Following Robert Boyd and Peter Richerson, Hwang and Bowles assume a degree of conformist cultural transmission, so

that the likelihood that an individual will adopt a particular preference varies not only with relative payoffs associated with the behaviors motivated by the preference but also with the prevalence of individuals with that preference in the population.

The second empirical regularity captured in their model of individual updating is that the presence and extent of incentives to contribute to a public project (or to engage in similar activities that benefit others) make the action (contribution) a less convincing signal of an individual's social preferences, resulting in observers interpreting some generous acts as merely self-interested. This is the key mechanism in the model of of Roland Benabou and Jean Tirole showing how incentives may crowd out pro-social behavior (Benabou and Tirole (2006)). Similarly, in his "Generous actors, selfish actions" paper, and in his subsequent work with Dufwenberg, Heidhues, Kirchsteiger, et al. (2011), Joel Sobel (2009) and his co authors provide not very restrictive conditions on individual utility functions such that "agents who care directly about the welfare and opportunities of others cannot be distinguished from selfish agents in market settings" (p.19). The reason is that for a class of utility functions admitting such other regarding preferences as inequality aversion, (paraphrasing the main theorem in their 2011 paper, p. 6) the "set of Walrasian equilibria of an economy [with other regarding preferences] coincides with the set of Walrasian equilibria of its corresponding ... economy [in which] agents care only about their own direct consumption." Thus the use of market-like incentives may make it impossible to infer generous or fair-minded behaviors from the observed actions of ones fellow citizens.

There are two reasons why the presence of an incentive may lead people to mistake a generous act – helping another at a cost to oneself -- for a self-interested one. The first is that the incentive provides a competing explanation of the generous act: "he did it for the money". The second is that incentives often induce individuals to shift from an ethical to a payoff maximizing frame (even relocating the neural activity to different regions of the brain); and knowing this, the presence of an incentive for an individual to help another may suggest to an observer that the action was self interested (Gneezy and Rustichini (2000b), Heyman and Ariely (2004), Irlenbusch and Sliwka (2005), Li, Xiao, Houser, et al. (2009)). The first "he did it for the money" reason depends on the magnitude of the incentive because in order to provide a convincing self-interested interpretation for the helping act the subsidy would have to exceed the cost of helping. The second reason --"when incentives are in force, everyone maximizes their payoffs" -- is categorical; it is simply the presence of the incentive that matters. Of course, an observer could make the opposite mistake, inferring that the generous act that was motivated entirely by an incentive, was done for ethical rather than payoff

maximizing reasons. In the model that follows the incentive is assumed on balance to degrade helping as the signal of a generous individual's type rather than motivating self interested individuals to act in ways that are mistakenly taken as signals of a generous type.

Taken together, these two assumptions imply that the extensive use of incentives may reduce the perceived population frequency of individuals with social preferences, leading (via the conformist learning effect) to an evolutionary disadvantage of generosity over self-interest in the preference-updating process. To show this Hwang and Bowles (2011b) adapt the model of endogenous preferences in Bowles (1998) and (2004) to study the effects of incentives on the preference-updating process. In terms of the state dependent model of the previous section, they study the effect of incentives on the equilibrium fraction of the population for whom λ_0 is positive and sufficient to motivate contribution to a public good.

Suppose there are two types: a Civic gives to the public good at a personal cost equal to g that may be partially offset by a subsidy s, while *Homo economicus* does not contribute and receives no subsidy. Both types update their traits by myopic best response, observing the material payoffs and public goods contribution of a sample of the population (they do not observe the utility of others) and a signal $\tilde{p}$ (possibly inaccurate when the planner implements a subsidy, $s > 0$) of the frequency of the Civics in the population, p, $\tilde{p} = \tilde{p}(s, p)$ which is decreasing in s. To capture the fact that the effect of the incentive on citizen's perception of the fraction of their fellow citizens who are Civics may depend on the mere presence of the incentive or on its extent, let

(7) $\quad \tilde{p} = p(1 + \mathbf{1}\{\boldsymbol{s} > 0\}\Lambda_c + s\Lambda_m)$

where as before the indicator $\mathbf{1}\{s > 0\} = 1$ if $s > 0$ and zero otherwise and $\Lambda_c \leq 0$ measures the categorical effect of the presence of an incentive on one's inference about another individual's type based on observing his or her contribution to the public good and $\Lambda_m \leq 0$ measures the marginal effect of the level of an incentive on one's inference. Note that when $p = 0$ or $s = 0$, $\tilde{p} = p$ so in the absence of the subsidy or when Civics are absent, the citizen's perception of the fraction of the population who are Civics is accurate.

The incentive has two offsetting effects on the distribution of types in the population, one intended and the other not: it raises the relative payoffs of the Civics, but it also reduces their apparent prevalence in the population. To see how this affects the equilibrium distribution of types in the population, suppose that individuals live forever but they periodically may switch their type. Denote the cultural fitness of trait i as r^i ($i = C, H$ for Civic and *Homo economicus*) defined as the expected number of replicas that each individual

bearing the trait will leave in the subsequent period. (If person k switches to j's type and j does not switch then k has left no replica and j has two replicas.) To capture the effect of socialization institutions on the evolution of preferences in this population the authors suppose that in any period some fraction γ of the H-types will be converted to a C-type. (Because it plays little role in what follows, Hwang and Bowles do not model the manner in which socialization institutions accomplish this, other than to assume that the process is not affected by the level of incentives). Then define $\alpha \in (0,1]$ as the relative weight of conformism rather than payoffs in the updating process, β as the weight of payoff differences relative to the socialization effect and π_C, π_H as the expected payoffs of the two types, so that the cultural fitness of the two traits can be written:

(8) $r^C = r_0 + \alpha\left(\tilde{p} - \frac{1}{2}\right) + (1-\alpha)\left[\beta(\pi_C - \pi_H) + \gamma\frac{1-p}{p}\right]$ and

(9) $r^H = r_0 + \alpha\left(\frac{1}{2} - \tilde{p}\right) + (1-\alpha)[\beta(\pi_H - \pi_C) - \gamma]$

The first term in both equations is the conformism effect, and it favors the Civics if it is perceived that they constitute more than half of the population. The second term is the net effect of socialization and payoff based updating. The socialization effect in equation (8) (the second term in the square brackets) is derived as follows: noting that poulation size is normalized to unity, each of the $1-p$ H types in the population has a γ probability of converting to C (shown in (9)) and thus appearing as $\gamma(1-p)$ replicas assigned to the p Cs in the population. The final expression in (8) is just the per C share of these socialized former H's.

From these cultural fitness equations one readily derives the familiar replicator equation for the movement of p over time:

(10) $dp/dt = p(1-p)(r^C - r^H)$

Introducing the costs of contributing to the public good and the subsidy and noting the payoff difference between the types $\pi_H - \pi_C$ is just $g - s$, the resulting stationary condition for an interior value of p (namely $r^C - r^H = 0$) is

(11) $\left(\tilde{p} - \frac{1}{2}\right)\frac{\alpha}{1-\alpha} = \beta(g-s) - \frac{\gamma}{2p}$

which requires that the conformist effect favoring the more common trait (the left hand side) offset the net effect of that trait's payoff disadvantages and the societal level socialization effects (the right hand side). Values of p satisfying (11) are termed the population's equilibrium preferences and denoted as $p^*(s)$. Figure 6 summarizes the relationship between the incentive and the prosocial action in the presence of endogenous preferences and Figure 7

illustrates the cultural equilibrium condition (11).

[Figure 6 and 7 here]

The solid lines in Figure 7 show the two sides of equation (11) – the conformist effect and the payoff plus socialization effects -- and their intersection, satisfying equation (11) when $s = 0$ and giving $p^*(0)$ that is, the equilibrium distribution of preferences in the absence of incentives. The dotted lines show the effect of the implementation of a subsidy. The intended effect is to reduce the payoff advantage of the H types (they do not receive the subsidy) shifting downward the payoff cum socialization function. A naive social planner, unaware of the conformist effect would thus expect the introduction of the incentive to increase the fraction of C's in the population to $p^N(s)$.

But the unintended effect of the subsidy is to reduce the perceived fraction of the population who are C's and thereby to diminish the conformist advantage of the C's. The downward shift in the conformist effect function thus partially offsets the payoff effect, with the resulting stationary distribution equal to $p^*(s)$. In the case of strong crowding out (not shown), the second effect would more than offset the first, resulting in a $p^*(s) < p^*(0)$.

The source of the non-separability between the socializations and incentives is clear if we return to equation (11) and consider the effect of an increase in s on the cultural fitness of the C types relative to the H types, evaluated at the status quo distribution of types in the population. This is just the vertical distance at $p^*(0)$ between the two functions that have been displaced by the introduction of the incentive (the dashed lines). Because this effect is the cultural fitness advantage of the C-types following the introduction of the incentive, Hwang and Bowles term it the evolutionary impact of the incentive, denoted by κ. (analogous to θ in the state dependent model of the previous section.) The direct effect of incentives on κ is just β, but as is clear from the following expression, there also is an indirect effect:

(12) $\frac{\Delta\kappa}{\Delta s}|_{p^*} = \beta + \frac{\alpha}{1-\alpha} p^*(\gamma, s)\left[\Lambda_{\mathrm{m}} + \frac{\Lambda_{\mathrm{c}}}{\Delta s}\right]$

where the left hand side means the change in the equilibrium condition associated with the change in s, for the given level of p, namely p^*. (As in the model of state-dependent preferences, we consider discrete changes here rather than simply differentiating (11), in this case because of the discontinuity of $\tilde{p}$ at $s = 0$ in the presence of categorical crowding).

The indirect effect will be negative in the case of crowding out, so the total effect of the incentive is less than the direct effect. The absolute size of the indirect effect (the second term

on the right of (12)) is (as expected) increasing in the extent of conformism in updating and in the (absolute magnitude of the) crowding parameters. Importantly, the negative indirect crowding out effect will be larger in absolute value, the greater is p^*. Because p^* varies positively with the socialization effect (γ), the total effect of the incentive is less, the more effective is a society's socialization institutions. The crowding effect will absent (separability will hold) if $\Lambda_m = \Lambda_c = 0$ in which case $\tilde{p} = p$, in which case there are no misperceptions of the fraction of C's in the population or $\gamma = 0$ in which case there are no C's to misperceive as self-interested, or $\alpha = 0$ in which case there is no conformism in updating so the mis-perceptions induced by the incentives have no effect.

We also have that the evolutionary impact of socialization institutions is

(13) $\frac{\Delta\kappa}{\Delta\gamma}|_{p^*} = \frac{1}{2p^*(\gamma,s)}$

which diminishes with greater use of incentives because (in the absence of strong crowding out) incentives raise p^*.

Thus where crowding out occurs incentives and socialization institutions are substitutes in the sense that the marginal effect of one on the evolutionary advantages of the civic minded types diminishes with the level of the other. We will return to the property of incentives and socialization as substitutes and the possibility of making them complements when we consider the policy implications of these models and the data to follow.

A summary of the two sources of non-separability – state dependence and endogeneity of preferences – and the mechanisms involved is provided in Table 2.

[Table 2 here]

The design of effective incentives in cases where separability may not hold requires a better understanding of the cognitive or affective effects of incentives that explain the categorical and marginal crowding out effects observed in experiments. We turn in the next three sections to the mechanisms that make preferences incentive-state-dependent, resulting in crowding out effects before considering (in section 8) the evidence for the adverse effects of incentives on preference-updating. (We consider crowding in --the case where incentives and social preferences are complements-- in section 9).

5. Bad news: Incentives provide information about the principal

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-information-about-the-incentive designer effect by the negative response to fines imposed by experimental 'investors' and 'trustees' in the Trust Game, a principal-agent experiment implemented by Ernst Fehr and Bettina Rockenbach.

German students in the role of "investor" were given the opportunity to transfer some amount to the other player, called the "trustee". This amount was then tripled by the experimenter. The trustee, knowing the investor's choice, could in turn "back-transfer" some (or all, or none) of this tripled amount, returning a benefit to the investor (Fehr and Rockenbach (2003)). When the investor transferred money to the trustee, he or she 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 known to the trustee and 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 back-transfers. Only one-third of the investors renounced the fine when it was available; their payoffs were 50 percent greater than the investors who threatened use of the fines.

The proximate causes of the negative impact of incentives in this case are suggested by evidence on the neural responses of the trustees in another Trust Game experiment (Li, et al. (2009)) As in the Fehr and Rockenbach experiment, the investor's threat of sanctions negatively affected back-transfers by trustees. To identify the proximate causes of this result, Jian 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 had threatened to sanction the trustee for insufficient back-transfers and an investor who had not threatened a sanction. Sanction threats de-activated the Ventromedial Prefrontal Cortex (VMPFC), a brain area whose activation was greater in trustees who made larger back-transfers experiment, as well as other brain areas thought to be involved in the processing of social rewards. The threat activated the parietal cortex, an area thought to be associated with cost-benefit analysis and

other self-interested optimizing processes. The interpretation by Li and his coauthors is that the sanctions induced a "perception shift" favoring a more self-interested response.

The signaling interpretation of counter-productive incentives in the Trust Game 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. The positive response to the investor's renunciation of the fine option is a categorical effect, analogous to the negative categorical effect of the use of incentives in the Irlenbusch and Ruchala experiment described above. The threat of the fine, however, conveyed a different message and diminished the trustee's reciprocity.

Similar cases of crowding out due to the "bad news" conveyed by the incentive are at work in experiments among student subject pools in Switzerland, U.S., Italy, France and Costa Rica (as well as Germany) and in a diverse set of games including Gift Exchange, Public Goods, and a charity giving setting similar to a Dictator Game. Costa Rican businessmen also responded negatively to the bad news that incentives conveyed. Table 3 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 [16, 18, 20, 21, 27]). Crowding out as the result of the "bad news" mechanism may be prevalent in Principal Agent settings and can be averted where the principle has a means of signaling trust or fairness (experiments [1-3]). Not surprisingly crowding out affects individuals who are intrinsically motivated or fair-minded (experiments [5-6]; for own payoff maximizers, it appears there is nothing to crowd out.

6. Moral disengagement: Incentives may suggest permissible behavior

In most situations people look for clues of appropriate behavior and incentives often provide them. In Table 4 we survey experiments in which this framing effect appears to have been at work. These experiments differ from those in Table 3, in which incentives were deployed by experimental subjects in the role of a principal interacting with an agent. Here incentives are implemented exogenously, that is by the experimenter, so that they provide no information about the intentions or beliefs of other experimental subjects. As can be seen from the table, incentives appear to affect moral disengagement not only among students but also (as we have seen) among poor Colombian villagers [12, 13] and top U.S. CEOs [16]. Moral disengagement was evident in the Ultimatum Game and the Common Pool Resource Game [11-14; 20-22] as well as in the games for which the bad news mechanism was at work

(table 3). In addition, this mechanism may be clearly recognized in settings of 1-player games (i.e. Dictator game or a performance Task) [30, 33].

Elizabeth Hoffman and her co author illustrated the framing 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" (Hoffman, McCabe, Shachat, et al. (1994)). The power of names has been confirmed in many (but not all) experiments since then (Zhong, Loewenstein, and Murnighan (2007)) but in some cases (Ellingsen, Johannesson, Möllerström, et al. (2011)) the framing effect appears to have altered subjects beliefs about the actions of others rather than their preferences.

But literally naming the game is not necessary for framing effects to occur. Incentives alone may provide powerful frames for the decision maker. A year before the first reality TV Survivor show, Andrew Schotter and his coauthors found that market-like competition for "survival" among subjects reduced their concern for fairness in an Ultimatum Game experiment (Schotter, Weiss, and Zapater (1996)). In this game Player 1 is given an endowment and asked to propose a division of it with Player 2. Player 2, knowing the size of the endowment, decides whether to accept or reject the division. If Player 2 accepts, then the proposed division is implemented. If Player 2 rejects, both players receive zero. As is commonly observed in the Ultimatum Game, Player 1 made quite generous offers and low offers were frequently rejected. But the experimenters told the subjects that those with lower earnings would be excluded from a second round of the game, Player 1 subjects offered less generous amounts to Player 2, and Player 2 accepted lower offers. The authors' interpretation was that: "...the competition inherent in markets...offers justifications for actions that, in isolation, would be unjustifiable."

While plausible, direct evidence for this "moral disengagement" explanation is lacking because the social preferences that apparently accounted for fair behavior in the non-survival condition of the experiment 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 detected.

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 US undergraduates (Barr, Wallace, Ensminger, et al. (2009), Henrich, Ensminger, McElreath, et al. (2010)). In the Dictator Game an experimental subject is assigned a sum of money and asked to allocate some all or none of it to a passive recipient. The Third Party Punishment Game is a Dictator Game with an active

onlooker (the third party) who observes the dictator's allocation. If the third party deems the dictator's allocation worthy of punishment he or she may then pay to impose a monetary fine on the dictator. One would expect that in the presence of a third party, the dictators would adjust their allocations upwards (compared to the two party standard Dictator Game) and thus avoid being fined. But this was not the case; 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 allocations were significantly (and in some cases substantially) lower. In Accra, for example, where 41 percent of the dictator's allocations resulted in fines by the third party, the allocations were 30 per cent lower in the Third Party Punishment Game than in the Dictator Game. The incentives provided by the fine did not induce higher allocations, 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 as we have seen does not preclude crowding out.)

Crowding out of ethical motives is suggested by the fact that the dictator's adherence to one of the world's religion (Islam or Christianity, including Russian Orthodoxy) raised allocations in the Dictator Game by 23 percent (compared to those unaffiliated with a world religion.). But in the Third Party Punishment Game, the estimated "religion effect" was reduced to just 7 percent of its value in the Dictator Game and it was not significantly different from zero. The presence of the incentive based on the fine appears to have defined the setting as one in which the moral teachings of these religions were not relevant. Consistent with a crowding out interpretation of these results, the negative effect on the dictator's allocations of his or her economic need (number of children, conditional on a given level of income and wealth) was substantial (and statistically significant) in the Third Party Punishment Game, but in the Dictator Game this "economic need effect" was an order of magnitude smaller and not significantly different from zero.

In the Accra sample (Barr (2004)) the dictator's allocation co-varied significantly with the frequency of attendance at church or mosque in the standard two party Dictator Game; but this large "religion effect" vanished in the Third Party Punishment Game. The incentives implicit in the Third Party Punishment Game appear to have substituted economic motivations for moral concerns. These experiments are also consistent with our model of state dependent preferences, in which crowding out operates via an effect of incentives on the behavior of those with pre-existing social preferences.

7. Control aversion: Incentives may compromise intrinsic motives and self-determination

Recent experiments by economists surveyed in Table 5 as well as non-experimental studies in economics (surveyed in Frey and Jegen (2001)) provide evidence for a third reason why social preferences may be state dependent in ways leading to crowding out. Table 5 does not include the original "over-justification" experiments done by psychologists (referred to in the introduction). Unlike the experiments by psychologists where incentives are typically implemented by the experimenter, economists often model strategic interactions in which the same apparently control averse reaction occurs, so these experiments could also fall under the "bad news" about the principal rubric presented in Table 3 (see [6,10]). Moreover, framing effects may result in moral disengagement in some of these experiments [24, 29, 33]. Crowding out effects of intrinsic motivation may be recognized in Ultimatum games [11, 12, 20] and games where the experimenter is the principal [18, 25, 30]. We think it is likely that in these and other cases more than one mechanism is at work.

Armin Falk and Michael Kosfeld used a principal-agent game to explore the idea that 'control aversion' based on the self-determination motive may be a reason why incentives sometimes degrade performance (Falk and Kosfeld (2006)). 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 simply whether or not to impose it.) The principal could infer that a self-interested agent would perform at the lower bound or, in the absence of the bound, at zero, and thus imposition of the bound would maximize the principal's payoffs.

But in the experiment agents provided a lower level of production when the principal imposed the bound. Apparently anticipating this response, fewer 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.

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 about them. In post-play interviews, most agents agreed with the statement that the imposition of the lower bound was a signal of distrust; and the principals who imposed the bound in fact had substantially lower expectations of the agents. The untrusting principals' attempts 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. Thus results in the Falk and Kosfeld experiment appear to be the result of both compromised self-determination and negative information about the incentive designer.

8. The economy produces people: Incentives alter how new preferences are learned

As in the Hwang and Bowles model introduced in section 4, incentives may also affect long-term change in motivations because they alter key aspects of how we acquire our motivations, influencing 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)).

Experiments of at most a few hours duration are unlikely to uncover the causal mechanisms involved in this process of durable preference change. 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. Acquiring new preferences (like a new accent) often takes place early in the life cycle and the learning process is strongly attenuated thereafter.

However, historical, anthropological, social psychological and other data (surveyed in Bowles (1998)) provide evidence for endogenous preferences, showing that economic structures affect parental child rearing values, personality traits rewarded by higher grades in school, and other developmental influences. Additional evidence that preferences are endogenous comes from the experimental studies of 15 small scale societies with extraordinarily varied economic structures, ranging from farming to hunting and gathering. In these studies cross subject pool comparisons showed a strong association between the nature of the diverse economic tasks required to secure a livelihood – participating in large

cooperative hunting teams in contrast to solitary work in forest slash and burn horticulture, for example -- and its members' experimentally measured generosity and fair-mindedness in the Ultimatum Game (Henrich, Boyd, Bowles, et al. (2005), Henrich, et al. (2010)).

Despite the limitations of experiments for the investigation of preference change, we survey in Table 6 a number of experiments that are consistent with durable learning effects of incentives. (We have placed all of the experiments consistent with preference endogeneity in this table; of course many of them also provide evidence of the mechanisms we have identified as affecting state-dependent preferences.) We take as evidence for this the fact that the apparent effect of incentives on preferences persists even when, in later stages of an experiment, incentives are withdrawn, suggesting that the prevalence of social preferences in a population may depend on exposure to incentives in the past, as in the Hwang and Bowles model.

An example follows. In the public goods experiment designed by Josef Falkinger, Fehr, Gächter, et al. (2000) an incentive mechanism induced subjects to contribute almost exactly the amount predicted for an own-material-payoff-maximizing individual, while in the absence of the incentive subjects contributed significantly more than would have been optimal for an own-material-payoff maximizing individual. But, consistent with a change in preferences due to exposure to incentives, in the absence of incentives, subjects who had previously experienced the incentive system contributed 26 per cent less than those who had never experienced it.

While the cultural diversity and variety of games appearing in Table 6 are substantial, and we think the preference learning effects that we have detected in these experiments are indeed at work, we do not yet have experiments capable of testing the mechanism underlying the models of the influence of incentives on the evolution of preferences proposed by Hwang and Bowles, Bar-Gill and Fersthman, and others

9. Crowding in

In section 2 we identified a number of cases in which crowding in may occur. For example the incentive may provide good news about the principal or it may lead to moral engagement rather than its opposite. In Table 7 we survey a number of studies that show this result. These experiments are of special interest to the social planner not only because they would ideally point the way to the design of policies which would make incentives and social preferences synergistic (that is complements) rather than substitutes, but also because it appears that crowding in occurs more often in games with more than 3 players (Public Goods

[42, 44, 45, 49, 50] and Common Pool Resource [14, 46] games) a common characteristic of public policy settings. In the penultimate and final section we will return to these questions when we consider the policy implication of non-separability.

Synergy between incentives and social preferences may explain why fines imposed on free riders by altruistic peers in a Public Goods Game induce higher levels of contribution in subsequent rounds of play (Fehr and Gächter (2000)). Of course crowding in need not have been involved; individuals might have simply best-responded to the anticipated loss in payoffs associated with low contributions. But more than this is at work. Consistent with the interpretation that incentives imposed by peers activate shame or other social preferences, purely verbal messages of disapproval have a substantial positive effect on free riders' subsequent contributions (Barr (2001), Masclet, Noussair, Tucker, et al. (2003)). When those who have contributed more than others are punished (as sometimes occurs, Herrmann, Thoni, and Gächter (2008a)), they subsequently contribute less, and costly retaliatory punishment sometimes results (Bowles and Gintis (2006), Carpenter, Bowles, Gintis, et al. (2009), Hopfensitz and Reuben (2009)). This appears to occur because the targets of the punishment feel hostility rather than shame.

There are also other mechanisms at work. The incentives and constraints typical of 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 (Bowles (2011)). This may explain 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 (those told this were not subject to the punishment.) They apparently wanted to be cooperative but wished even more to avoid being exploited by defectors. According to this interpretation, the fine imposed by the experimenter on any free riding liberated the individual to act pro-socially without fearing being exploited by less cooperative players. The respondents may have exhibited what Iris Bohnet and her co authors call "betrayal aversion," which was attenuated by knowing that betrayal would be punished by a third party (Bohnet, Greig, Herrmann, et al. (2008)).

Market incentives may also favor the endogenous evolution of social preferences. In two sets of experiments in small-scale societies in Africa, Asia and Latin America (Henrich, et al. (2005), Henrich, et al. (2010)), individuals from the more market-integrated societies gave more in the Ultimatum Game. The authors conjecture that this may be due to the fact

that more market exposed subjects had the experience of mutually beneficial exchanges with strangers, much like in the anonymous experimental settings. A very different piece of evidence consistent with this interpretation is that subjects who were exposed to unobtrusive priming with words relating to markets and exchange prior to playing a Trust Game were more likely to trust their partner than were subjects exposed to primes unrelated to markets (Al-Ubaydli, Houser, Nye, et al. (2011)).

A distinct mechanism underlying crowding in was apparently at work in a public goods experiment by Pietro Vertova and Roberto Galbiati. Consistent with the Cardenas experiment described in section 2, 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. A stronger monetary incentive did not result in an increase in contributions. The strong monetary incentive also had no effect on behavior in the absence of the stated obligation (Vertova and Galbiati (2010)). The authors' interpretation (like that of Cardenas) is that the explicit incentives enhanced the salience of the stated obligation. In our taxonomy it is a case of categorical crowding in (See also Galbiati and Vertova (2008)).

10. The lab and the street: Can one generalize from experimental evidence?

The experimental evidence for non-separability would not be very interesting if it 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 a concern in any empirical study (Falk and Heckman (2009)) and is often unwarranted (Levitt and List (2007)). Consider, for example, the Dictator Game: typically more than 60% of the dictators allocate a positive sum to the recipient, and the average given 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 an anonymous passerby, or that the same subjects would offer a fifth of the bills in their wallet to a homeless person asking for help. Another example: while pro-social behavior in an experiment by Benz and Meier (2008) was correlated with non-experimental behavior, subjects who reported that they had never given to a charity allocated 65 percent of their endowment to a named charity in a lab experiment.

A possible explanation of these discrepancies between experimental and real world behavior is that 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 an exception to this context-dependent aspect of

individual behavior. External validity concerns arise from four aspects of human behavioral experiments that do not arise in most well-designed natural science experiments. First, experimental subjects typically know they are under an unknown researcher's microscope, possibly inducing different behaviors than would occur under total anonymity or under the scrutiny of neighbors, family or workmates. Second, experimental interactions with other subjects are typically anonymous and without opportunities for ongoing face to face communication, unlike many social interactions of interest to economists and policy makers. 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 or the Public Goods Game that give scope for ethical or other-regarding behavior that may be absent in competitive markets and other important real world settings (Sobel (2010)).

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 be less motivated by social preferences subject to crowding out; or knowing that the topic of the experiment was cooperation the subjects might be atypically civic minded.

We can do more than speculate about these problems. Nicole Baran and her coauthors asked if University of Chicago Graduate School of Business students who were more reciprocal in the Trust Game (those who as trustees most generously reciprocated large transfers by the investor) were also those most likely to contribute to the University upon graduation. They were (Baran, Sapienza, and Zingales (2010)). Fehr and Lorenz Goette

found that in a group of bicycle messenger workers in Zurich, those who exhibited loss aversion in a laboratory experiment exploring the subjects' preferences over lotteries also exhibited loss aversion when faced with real-life wage rate changes (Fehr and Goette (2007)). Dean Karlan (2005) implemented a Trust Game among Peruvians participating in a micro-credit program; those who were least trustworthy (transferred less back to the "investor") in the experiment were less likely to repay their real world loans. Alain Cohn and his co authors (Cohn, Fehr, and Goette (2011)) found that reciprocators in the lab (measured by play in a sequential PD game) responded positively to a randomly awarded fixed wage increase in their work, while those who played the sequential PD in a payoff maximizing way did not respond to the wage increase.

Among the Japanese shrimp fishermen that Jeffrey Carpenter and Erika Seki studied, those who contributed more in a public goods experiment were more likely to be members of cooperatives that shared costs and catch among many boats than to fish under the usual private boat arrangements (Carpenter and Seki (2010)). A similar pattern was found among fishermen in the Brazilian north east, where some fish offshore in large crews whose success depends on cooperation and coordination, while those exploiting inland waters fish singly. The ocean fishers were significantly more generous (in Public Goods, Ultimatum and Dictator Games) than the inland fishers (Leibbrandt, Gneezy, and List (2010)).

A better test of the external validity of experiments would include a behavior-based measure of how cooperative the individuals were, not simply whether they took part in a cooperation-sensitive production process. The Brazilian fishers provide just such a test. Shrimp are caught in large plastic bucket-like contraptions; holes are cut in the bottom of the traps to allow the immature shrimp to escape, thereby preserving the stock for future catches. The fishermen thus face a real world social dilemma: the present value of expected income of each would be greatest if they cut only small holes in their own traps while others cut large holes in their. Small trap holes are a form of defection, and just as in the Public Goods Game it is the dominant strategy for a self-interested individual. But a shrimper might resist the temptation to defect if he were both public spirited towards the other fishers and sufficiently patient to value the future opportunities that they would lose were he to use traps with smaller holes. Fehr and Andreas Leibbrandt implemented both a Public Goods Game and an experimental measure of impatience with the shrimpers. They found that both patience and cooperativeness in the game predicted larger trap holes (Fehr and Leibbrandt (2011)). The effects, controlling for a large number of other possible influences on hole size, were substantial. A shrimper whose experimentally measured patience and cooperativeness is a standard deviation greater than the mean is predicted to cut holes in his traps that are half a standard deviation larger than the mean.

Additional evidence of external validity comes from a set of experiments and field studies with 49 groups of herders of the Bale Oromo people in Ethiopia who were engaged in forest commons management. Devesh Rustagi and his coauthors implemented public goods experiments with a total of 679 herders. They also studied the success of the herders' cooperative forest projects. The most common behavioral type in the experiments, constituting a bit more than a third of the subjects, were "conditional cooperators" who responded positively to higher contributions by others. Controlling for a large number of other influences on the success of the forest projects, the authors found that groups with more

conditional cooperators were more successful, in terms of number of new trees planted, than groups with fewer conditional cooperators. This was in part because members of groups with more conditional cooperators spent significantly more time monitoring the use of the forest by others. As in the case of the Brazilian shrimpers, the effects of group composition were large. A 10% increase in the fraction of experimentally identified conditional cooperators in a group was associated with an increase in trees planted or time spent monitoring by members of the group of about 3% (Rustagi, Engel, and Kosfeld (2010)).

The available evidence suggests that students volunteering for experiments are not more pro-social in their orientations than other students (Falk, Meier, and Zehnder (2011)); nor are student subjects more pro-social than non-students, indeed the reverse seems to be the case. (Fehr and List (2004), List (2004), Cardenas (2005), Carpenter, Verhoogen, and Burks (2005), Bellemare, Kröger, and Van Soest (2008), Carpenter, Connolly, and Myers (2008), Burks, Carpenter, and Goette (2009), Baran, et al. (2010), Cleave, Nikiforakis, and Slonim (2010), Cardenas (2011), Falk, et al. (2011) and see Supplementary online material for a description of these studies.)

While warranting caution in generalizing the details of experimental behavior to the real world, none of the external validity concerns is sufficient to dismiss the experimental evidence that social preferences are important behavioral motivations and that these preferences may be affected by explicit incentives. This is especially the case when experimental subjects exhibit motives such as reciprocity, generosity and trust that allow a consistent explanation of otherwise anomalous real world examples of crowding in or out, such as those mentioned at the outset.

11. Optimal incentives for the sophisticated social planner

There are multiple plausible interpretations of the mechanisms underlying non-separability in the experiments we have presented, as is clear from the substantial size of the intersections among the hypothesized crowding out mechanisms that is evident in Figure 2. It would nonetheless be difficult, in light of these data, to sustain the implicit separability assumption adopted in many economic models.

A sophisticated social planner (or mechanism designer) – one who knows that the separability assumption is likely to be violated – faces a challenge that has yet to be addressed in the public economics literature: how to design optimal taxes, fines, or subsidies when the preferences that will determine citizen's responses depend on the incentives deployed. Thus, the designer must consider the effects – whether state-dependent or endogenous – of the

instruments under consideration on individuals' social preferences and evaluate alternative policies on the basis of the resulting joint equilibrium of these preferences and economic allocations.

The problem facing the planner is quite a bit more difficult than the one we faced writing this paper. We studied the effects of incentives in experiments and natural settings and then sought *ex post* to determine the kinds of non separability – categorical or marginal crowding out or in – that might explain the results. The planner, however, must determine, *ex ante* whether the separability assumption is likely to be violated, and if so, how. The challenge is even greater because the nature and extent of non-separability itself is not given but (as we will see) may be influenced by the overall policy package of which the incentives are a part.

We begin with the more modest way of addressing the planner's problem and consider the nature and degree of the indirect effects of incentives on social preferences (that is, the signs and the size of the crowding parameters $\lambda_c, \lambda_m, \Lambda_c$ and Λ_m) as exogenously given and simply determine the optimal level or mix of incentives taking account of their effects on preferences (Fershtman and Heifetz (2006), Heifetz, Segev, and Talley (2007), Bowles and Hwang (2008), Hwang and Bowles (2011a)).

Here, two results may guide the social planner. The first is that in the presence of crowding out, incentives and social preferences are substitutes, so the deleterious indirect effect of incentives will be least where individual social preferences are modest or nonexistent (as will be the case in the endogenous preference model if there are few or no public spirited citizens or in the state dependent model where the citizen's baseline social preferences are modest or zero). Societies in which social preferences are more prevalent not only may be able to afford less use of incentives but will find them less effective (when both direct and indirect effects are accounted for) than would be the case in a less civic minded culture. By the symmetry of the definition of substitutes (see equations 6 and 13) in the presence of crowding out, policies to enhance social preferences (that is raising γ or λ_0) will be more effective in promoting contributions the public good where incentives are little used.

In a cultural-institutional dynamic setting where economic incentives and socialization practices to promote civic mindedness are adopted as alternative measures to enhance public goods provision, this substitutability property of incentives and social preferences may support at least two evolutionarily stable equilibria. In one, extensive use of incentives is coupled with relatively low levels of civic mindedness in the population. In this state there is little incentive to inculcate social preferences, the effect of which would be modest given the

crowding phenomenon. In the other cultural-institutional equilibrium a social planner serving a civic minded population makes more modest use of incentives due to their limited effectiveness, once their crowding out effects are accounted for.

The second result for the social planner takes us back to Titmuss and others who concluded that if incentives crowd out social preferences then incentives will be overused by a naïve planner who is unaware of the effects of incentives on preferences. As a result, in these cases the sophisticated planner would either not use incentives, or would use them less than would the naïve planner. But the prescription that incentives are overused does not follow from the (correct) observation that crowding out occurs: it is readily shown that when crowding out occurs the sophisticated planner may make either greater or lesser use of explicit incentives than would her naïve counterpart (Bowles and Hwang (2008), Hwang and Bowles (2011a)).

The sophisticated planner may make greater use of incentives when incentives crowd out social preferences is that if incentives work less well than would be the case under separability, then there are two offsetting influences on their optimal use. The one that forms the basis of the Titmuss critique is that crowding out reduces the marginal effect of the subsidy on the target's behavior; and if this were the only effect Titmuss would be right. But there is a second often overlooked effect. Because the incentive is less effective (either categorically or marginally), the under provision of the public good will be exacerbated (compared to what would occur were crowding out absent) and if the benefits of the public good are concave in the amount provided the marginal benefit of altering the target's behavior is therefore correspondingly greater.

The intuition is transparent: the doctor who discovers that a treatment he has been prescribing is less effective than he thought may opt for stronger doses rather than weaker or for abandoning the treatment. As long as there are diminishing marginal returns to the public good and crowding out is categorical (and not too large) the naïve social planner will make too little use of the incentive. The reason is that in this case crowding does not change the marginal effect of the incentive on the citizens' contribution level; but the reduction in the public good resulting from crowding means that the marginal benefits to increasing its supply rise. (If categorical crowding is sufficiently large the naïve planner will over-use the incentive because the sophisticated planner will choose no incentive at all in this case.) But the sophisticated planner may make greater use of incentives even when only marginal crowding out occurs, if the benefit function is sufficiently concave.

A less modest approach to the design of appropriate incentives where separability may not hold is to recognize that the extent of the non-separability problem (that is, the magnitudes of the crowding parameters in the models of section 3 and 4, namely λ_c, λ_m, Λ_c, and Λ_m) is not exogenous, but can be affected by the nature of the incentives and the manner in which they are deployed. Designing policies that can convert incentives from being substitutes for social preferences to being their complements, however, requires an understanding of why crowding out occurs.

The most plausible explanation for the failure of the separability assumption is 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 (Cooley (1902), Leung and Martin (2003), Akerlof and Kranton (2010), Bloom (2010)). We refer to the second – the being or becoming motives – as constitutive. Incentives addressed to our acquisitive desires sometimes appear to dampen or impede the pursuit of our constitutive aspirations. Among the reasons, we have seen, are that in addition to affecting the costs and benefits of an action, incentives also provide information about the person imposing the incentive, suggest appropriate behavior by framing decision situations, may compromise the target's sense of autonomy, and alter the environments in which we learn new preferences.

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 of violence they might be if their political leaders were to compromise on contested issues between the groups. Those who regarded their group's claims (on Jerusalem, for example) as reflecting "sacred values" (about half in each of the three groups) expressed far greater anger, disgust and support for violence if the compromise were accompanied by a monetary compensation for their own group than if no compensation were offered (Ginges, Atran, Medin, et al. (2007)). Similar results were fund in a survey of the willingness of Swiss citizens to accept environmental hazards (Frey and Oberholzer-Gee (1997)). (For a discussion on environmental motivation and crowding effects see Frey and Stutzer (2008).)

The importance of constitutive rather than acquisitive motives may be at work in the negative response to incentives that convey adverse information about the individual imposing the incentives. Recall that in the Trust Game implemented by Fehr and Rockenbach (2003) the investor's threat to fine the trustee if the back transfer was not sufficient had the effect of reducing the level of reciprocity of the trustee: conditional on the investor's transfer to the trustee, back-transfers were less under the fine condition. This was especially the case

when it appeared that the intent of the fine was to induce the trustee to grant most of the joint surplus to the investor. Where the investor announced modest levels of desired returns such that the investor and the trustee would both substantially share in the joint surplus, the use of the fines reduced back-transfers by an insignificant amount. But where the announced desired back-transfer would have allowed the investor to capture most of the surplus had the trustee complied, the reduction in back-transfers was 38 percent. It appears that the use of the fine in these conditions signaled the unfair intent of the investor, rather than simply his distrust of the trustee.

The fact that in this latter case incentives appear to have revealed 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 (Fehr and Gächter (2000) and (2002a), Masclet, et al. (2003)). One treatment in these public goods experiments is particularly revealing: group membership is shuffled after each period so that in subsequent periods a punisher will not be in the same group with the target of his or her punishment, and thus cannot benefit from the target's response. Punishment in this case is an altruistic act as it benefits others at the expense of the punisher and hence it cannot be interpreted as a signal of unfair intent. In this setting there is a strong positive response by low contributors.

Although there is no direct evidence, a plausible explanation of the effectiveness of incentives in this case is that when punished by a peer who had nothing to gain by doing so, those who have contributed less than others interpret the punishment as a signal of public-spirited social disapproval by fellow group members seeking to uphold a social norm and willing to sacrifice payoffs to do so. As a result, targeted free riders and even free riders who escaped punishment feel shame, which they redress by subsequently contributing more. In this case the incentive (prospect of peer imposed fines) has crowded in social preferences. These experiments illustrate the opposite of the “bad news about the principal” results in section 5. The principals here are the peers who punish free riding fellow group members; and the positive response to the fines in this case may reflect the fact that the willingness to pay to punish defectors with no expectation of personal gain is good news about the person implementing the incentive.

Consistent with the interpretation that crowding out does not follow from the use of incentives per se, but rather from the meaning that the incentives convey to the participants is

an extension of the "control aversion" experiment of Falk and Kosfeld (2006) described in section 5. Schnedler and Vadovic (2011) found that when agents themselves implemented controls (rather than the principal) the negative response did not occur. A large number of experiments have found positive effects of incentives imposed by the decision of the targets of the incentives rather than by the experimenter or by a principal (Cardenas (2005), Tyran and Feld (2006), Kroll, Cherry, and Shogren (2007), Ertan, Page, and Putterman (2009), Kosfeld, Okada, and Riedl (2009), Mellizo, Carpenter, and Matthews (2011), Sutter, Haigner, and Kocher (2011)).

John Stuart Mill (whose definition of the restrictive boundaries of our discipline we mentioned at the outset) and economists since have recognized that the purposes of individual economic action are constitutive as well as acquisitive (Akerlof and Kranton (2010)). But what some have missed is that our acquisitive and constitutive motivations may not be separable.

Some of the founders of economics knew this. Jeremy Bentham's *Introduction to the Principles of Morals and Legislation* (1789), is arguably the first text in what we now call public economics. In it he explained how proper incentives should harness self-interested objectives for public ends by making "it each man's *interest* to observe ... that conduct which it is his *duty* to observe." In other words, make sure that doing his duty is incentive compatible.

But he also understood the constitutive side of action and the need to design incentives that are complements of the moral sentiments rather than substitutes:

> A punishment may be said to be …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 (1970 [1789]): p.26).

Few economists followed Bentham in this. An exception is Albert Hirschman, who noted that economists seek

> to deal with unethical or antisocial behavior by raising the cost of that behavior rather than proclaiming standards and imposing prohibitions and sanctions. The reason is probably that they think of citizens as consumers with unchanging or arbitrarily changing tastes in matters civic as well as commodity-related behavior. . . A principal purpose of publicly proclaimed laws and regulations is to stigmatize antisocial behavior and thereby to influence citizens' values and behavioral codes. (Hirschman (1985): p.10)

The fact that punishments are "moral lessons" that "stigmatize antisocial behavior" 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 at 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 the opposite effect: in two weeks it resulted in a 94 percent decline in their use and appeared to crowd in social preferences (Rosenthal (2008)).

The contrast is instructive. In the Haifa case, the experimenters (respecting standard experimental protocols) provided no justification for the introduction of the fine on the tardy parents. Moreover the parents' occasional lateness could have occurred for reasons beyond their control, rather than as the result of a deliberate disregard for the inconvenience it caused the teachers. Finally, lateness was not so common as to be widely broadcast to the other parents. By contrast, the introduction of the Irish plastic bag tax was preceded by a substantial publicity campaign, and the use of the bags required a deliberate choice made in a highly public condition. In the Irish case, as in the experiment by Vertova and Galbiati (2010)) mentioned in section 9 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 disposal of the bags.

The same message comes from a voting study. In Switzerland the removal of a negligible fine for not voting significantly reduced voting turnout; but a considerable reduction in the cost of voting (by allowing balloting by mail) had no effect on turnout. The implication is that the fine for not voting encouraged turnout not as an incentive (by affecting the costs of not voting) but rather as a message of the importance of one's civic duty (Funk (2007)).

The fact that fines often work more as messages than as incentives poses a problem for the sophisticated planner because the same intervention may bear radically different messages in different cultures. Bohnet and her co authors implemented a Trust Game in which in one treatment the investor had the option of reducing the payoffs of trustees who betrayed their trust (Bohnet, Herrmann, Al-Ississ, et al. (2010)). Compared to the treatment in which this so-called "revenge" option was not available, when they had the revenge option a substantially larger fraction of Saudi investors trusted their partner, while a substantially smaller fraction of American investors trusted. Making trust more incentive compatible thus had diametrically opposed effects in the two cultures.

12. Conclusion: Are incentives to blame?

Is there a simple lesson for public policy? We think there is. Titmuss was right that incentives sometimes crowd out non-economic motives, and this may degrade economic performance. But Titmuss and the literature that followed him targeted incentives *per se* as the cause of crowding out and recommended a reduced role for incentives in the governance of economic interactions.

Both the diagnosis and the policy implication are wrong. Crowding out, as we have seen, may require greater, not lesser use of incentives. And perhaps more important: fines, subsidies, and other monetary incentives *per se* may not be the culprit. What accounts for crowding out, we believe, is the meaning of the fines or subsidies to the target of the incentives; and this depends on the social relationships among the actors, the information the incentive provides, and the pre-existing normative frameworks of the actors. This is the message of the contrast between the Irish grocery bag tax and the Haifa fines for tardiness, along with the fact that fines imposed on low contributors by peers in Public Goods Games have positive effects while fines imposed by principals on agents sometimes backfire. In addition, incentives chosen by agents (for example by majority rule), may have a more positive effect on individual performance than if they are imposed (Mellizo, et al. (2011)).

Fines deployed either to exploit or to control the target (or that give this appearance or that have this effect) are likely to be less effective than they would under separability and may even be counterproductive. The reason, we think, is that they activate the target's desire to constitute himself or herself as a dignified and autonomous individual who is treated fairly by others. It is this constitutive motive that sometimes trumps the acquisitiveness tapped by the incentive, and that leads to a contrary response. The same incentives deployed by individuals who do not stand to benefit personally, and that are intended to foster pro-social behavior are more likely to be complements rather than substitutes for social preferences, crowding them in rather than out. They do this by activating rather than diminishing the target's constitutive motives such as the desire to be treated fairly and to treat others fairly, to be a good member of a community, and the feeling of shame when others regard one as having failed in this.

Present experimental and other evidence give insufficient guidance to the planner who wishes to know ex ante, the effects of the incentives he is considering implementing. But on the basis of what we do know a good rule might be the following: The policy package of which the incentives are part should let the target understand that the desired modification in her actions will serve to implement an outcome that is socially beneficial so that that the target is more likely to endorse the purpose of the incentive, rather than being offended by it

as either unjust or a threat to her autonomy or in some other way reflecting badly on the intentions of the planner.

Figures

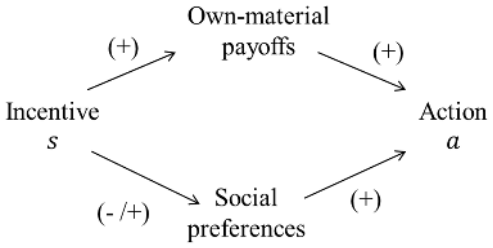

Figure 1. Crowding effects of incentives: The direct and indirect effects of incentives on contribution to a public good (a) The effect of an incentive (s) on social preferences may be either to reduce their behavioral salience for the action (social preferences are state-dependent) or to affect the manner in which preferences are updated, thereby altering the individual's social preferences (endogenous social preferences). Crowding out occurs when the effect of an incentive on social preferences is negative (assuming that the effect of social preferences on the action is positive, as shown). Crowding in (the opposite) also occurs.

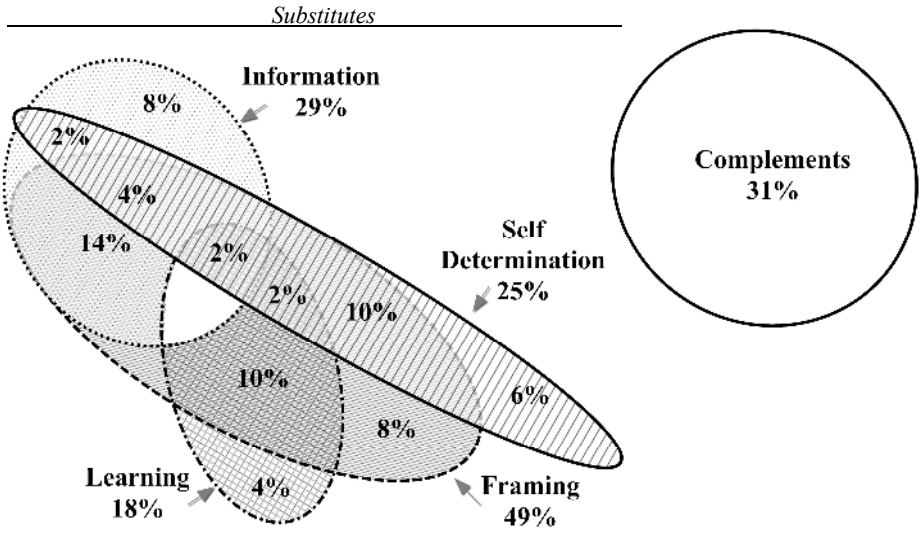

Figure 2. Summary of experimental evidence on the four crowding out mechanisms and crowding in. In the figure on the left the mechanisms accounting for crowding out are shown. The intersections show cases in which more than one mechanism may be involved. For example 14% of the experiments are consistent with both the framing and information about the incentive designer mechanisms. The circle in the upper right refers to crowding in (we have separated out the mechanisms in this case.) The numbers indicate the percentage of the total of 50 studies that exhibit the mechanisms indicated. There are no studies in the intersections that are blank.

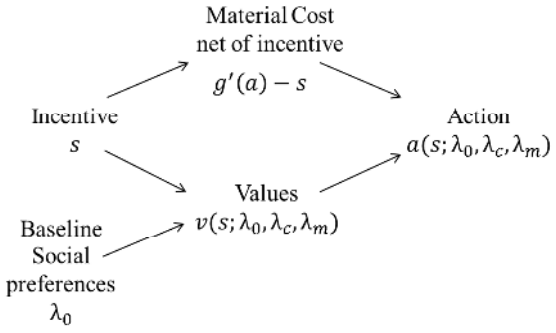

Figure 3. Crowding effects of incentives for an individual with state-dependent preferences. Baseline social preferences are the individual's non-material motivations to contribute to the public good in the absence of an incentive. Incentives reduce the net cost of contributing to the public good; but unless $\lambda_c = 0 = \lambda_m$ (separability) or $\lambda_0 = 0$ (no social preferences to crowd out) they also affect the motivational salience of the individual's social preferences.

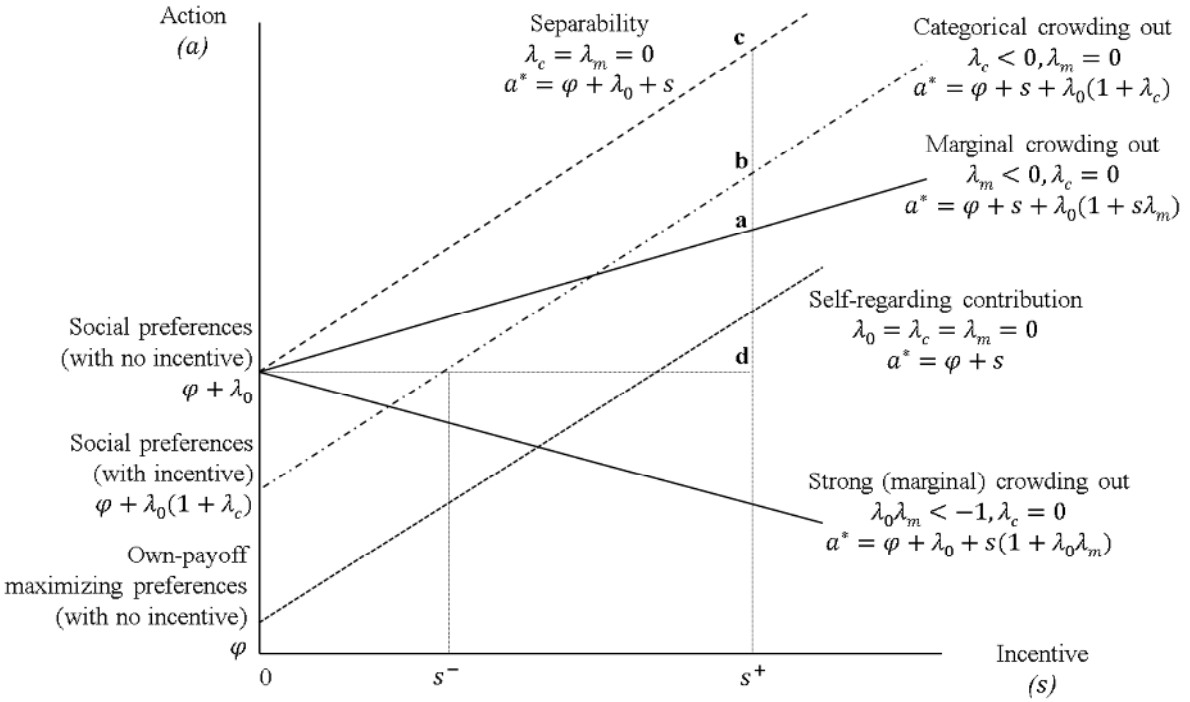

Figure 4. The sophisticated planner's implementation technology: **Citizen's contribution to the public good with state-dependent social preferences.** Under separability (top dashed line) incentives and social preferences are additive. Under strong marginal crowding out the use of the incentive is counterproductive (i.e. reduces contributions). Under categorical crowding out (dot-dashed line), incentives are also counterproductive for sufficiently small $s < s^-$.

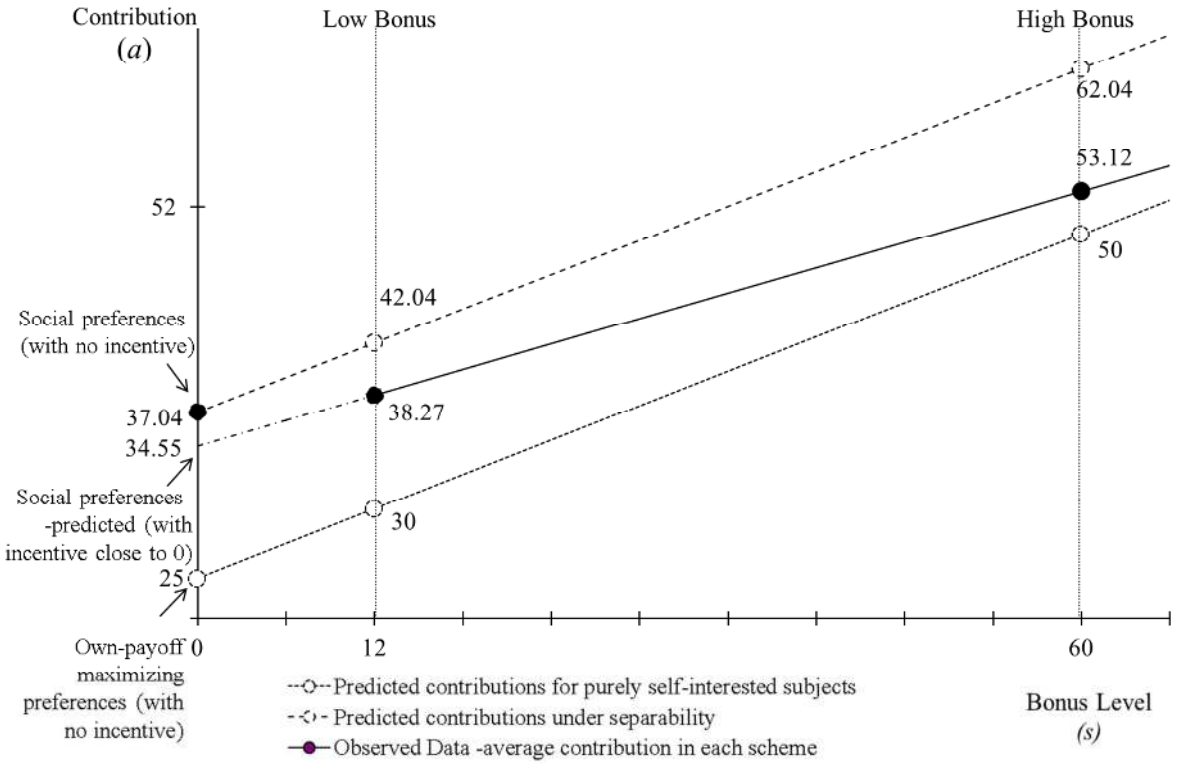

Figure 5. Categorical and marginal crowding out. Source: calculated from Irlenbusch and Ruchala (2008). See text. The experimental design is an adapted Public Goods Game comparing two team-based compensation schemes without and with a relative bonus.

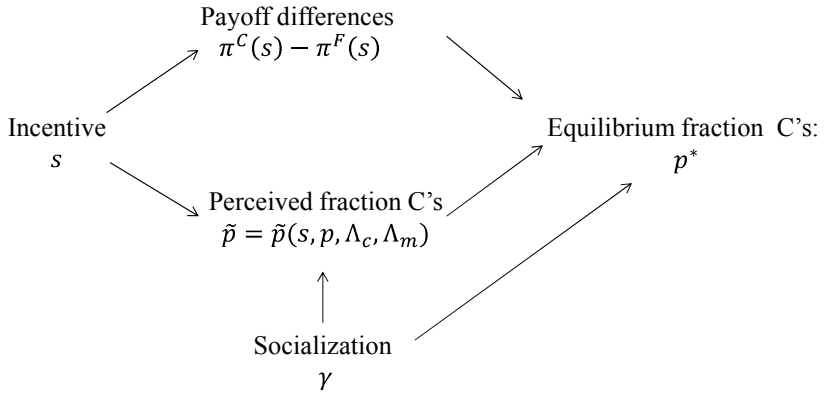

Figure 6. Crowding effects of incentives for a population with endogenous preferences. Incentives raise the relative payoff of those contributing to the public good, supporting a larger fraction of civic-minded citizens; but unless $\Lambda_c=0=\Lambda_m$ (separability) or $\underline{p} = 0$ in which case $p^* = 0$ (no civic minded citizens in the absence of a subsidy) a subsidy also alters the preference updating process by reducing the perceived fraction of civic-minded citizens $\tilde{p}$.

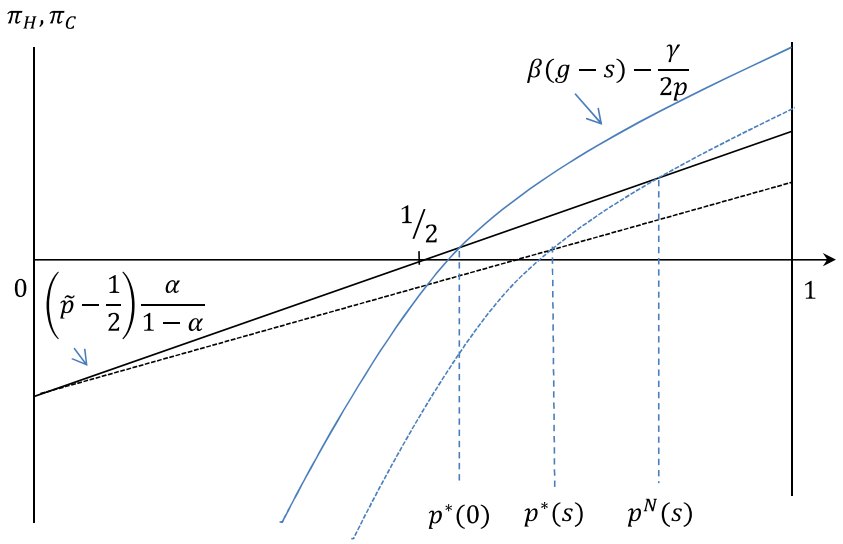

Figure 7. Incentives and equilibrium preferences. The figure (solid lines) shows the determination of fraction of citizens with social preferences in a cultural equilibrium under the influence of payoff-based and conformist updating, namely p^* when $s = 0$.The subsidy (dotted lines) reduces the payoff difference between *Homo economicus* and Civics, and in the absence of the effect on the perceived frequency of conformism in the population, the fraction of C's in the population would increase from $p^*(0)$ to $p^N(s)$. However, the reduction in the conformism effect partially offsets this. The resulting equilibrium outcome is $p^*(s)$. Source: Hwang and Bowles (2011b)

Source and Characterization (modeled in section §)	**Mechanisms**	**Description (§: section with empirical evidence relevant to this mechanism)**
State-dependent preferences Incentives affect the *behavioral salience of an individual's social preferences*, §3	**Information "bad news"**	Incentive signals the designer's type or beliefs about the target or the nature of the targeted task, and may convey illegitimate pursuit of self-interest by principal. §5
	Framing "moral disengagement"	Incentive signals the type of situation and hence appropriate behavior for the target, and may activate own payoff-maximizing modes of thought. §6
	Self-determination "control aversion"	Incentive affects target's sense of autonomy, and may signal unacceptable control and motivate resistance. §7
Endogenous preferences Incentives affect the *environment in which preferences are learned* and therefore the stationary *distribution of preference types in the population* (i.e. the fraction of population with social preferences), §4	**Conformist preference-updating**	Incentives reduce the perceived population fraction of social preference types. The extent to which a society relies on economic incentives – as opposed to other kinds of motivations and controls – will affect how people learn new preferences that may persist over long periods. §8

Table 1. Economic incentives and social preferences: Endogenous and state-dependent effects and mechanisms. As a result of the mechanisms listed incentives and social preferences may be either complements (crowding in) or substitutes (crowding out). In the conclusion we consider cases in which the degree of endogenous or state-dependent non-separability is subject to public policy because the crowding parameters $\lambda_m, \lambda_c, \Lambda_m$ and/or Λ_c may themselves be affected by incentives. Additional mechanisms for endogenous crowding out are provided in Bowles (2004).

Preferences	**State dependent**	**Endogenous**
Exogenous determinant of social preferences	Individual baseline values λ_0	Population level socialization effect γ
Crowding mechanism	Salience of values $v = \lambda_0(1 + \mathbf{1}\{s > 0\}\lambda_c + s\lambda_m)$	Perceived fraction C's $\tilde{p} = p(1 + \Lambda_c + s\Lambda_m)$
Intended target of the incentive	Individual best response $a^*(s, g(a), \lambda_0, \lambda_c, \lambda_m)$	Fraction of Cs in population $p^*(\mathrm{s}, \mathrm{g}, \gamma, \Lambda_c\Lambda_m)$
Separability	$\lambda_0\left(\frac{\lambda_c}{\Delta \mathrm{s}} + s\lambda_m\right) = 0$	$\frac{\alpha}{1-\alpha} p^*(\gamma, s)[\Lambda_m + \Lambda_c/\Delta \mathrm{s}] = 0$
Sufficient Conditions: separability	$\lambda_0 = 0$ or $\lambda_c = \lambda_m = 0$	$\gamma = 0, \alpha = 0$ or $\Lambda_c = \Lambda_m = 0$
Necessary Conditions: crowding out (in)	$\lambda_0 > 0$, λ_c or $\lambda_m < (>)0$	$\gamma > 0$, $\alpha > 0$ Λ_c or $\Lambda_m < (>)0$

Table 2 Separability and crowding when social preferences are state dependent or endogenous. In both models the citizens may bear a cost ($g(a)$ or g) in order to contribute to a public good where a subsidy, s, may partially offset the cost. In the endogenous preference model those who contribute are C's. Additional notation: λ_m , λ_c and Λ_m ,Λ_c are the marginal and categorical crowding parameters (in the state-dependent and endogenous cases, respectively) and α is the relative importance of conformism in the endogenous preferences model..

Tables 3 to 7.
Note: The bold entries in the comments column -- **I, F, S, E** and **C** -- indicate that the experiment in question could also have been included in tables 3 (**I**nformation about the principal) 4 (**F**raming) 5 (**S**elf-determination) 6 (**E**ndogenous preferences) or 7 (**C**omplementary relations between incentives and social preferences). All the papers but those marked with an * are published or forthcoming in a publication. The entries for each table are organized as follows: First, those studies that are published in a journal, ordered by year and first author. Second, working papers, ordered by year and first author.

Table 3. Bad news: Incentives provide information about the person who implements the incentive (I)

Citation	Subjects (number)	Games or activities	Institutional environments (treatments)	Results relevant to separability	Comment (quotes are from the cited paper)
[01] Fehr and Rockenbach (2003)	German students (238)	Trust Game	• Optional punishment as an incentive contract (i.e. a fine if less than the desired back-transfer amount is returned). The level of the fine is fixed by the experimenter and the only choice of the investor is whether to impose the fine or not	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). 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**
[02] Fehr and List (2004)	Costa Rican CEOs (126) and 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).
[03] Borges and Irlenbusch (2007)	German Students (179)	Buyer - Seller Game	• 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**
[04] Fehr and Schmidt (2007)	German Students (70)	Gift-Exchange Game	• Two internal forms of enforcement: The principal (employer) 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.

Table 3. Bad news: Incentives provide information about the person who implements the incentive (I) (Continued…)

Citation	Subjects (number)	Games or activities	Institutional environments (treatments)	Results relevant to separability	Comment (quotes are from the cited paper)
[05] Fehr, et al. (2007)	German students (130)	Gift-Exchange Game	• Three internal forms of enforcement: The principal can choose to rely on - a trust (pure fixed payment) 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	Under the unenforceable bonus contract subjects contribute more than the payoff maximizing Nash equilibrium, outperforming the enforceable incentive contract (fine). The results are the same independently of the framing.	"Bonus contracts that offer a voluntary and unenforceable bonus for satisfactory performance provide powerful incentives and are superior to explicit incentive contracts when there are some fair-minded players".
[06] Dickenson and Villeval (2008)	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 partner treatment, when employer payoffs depend on employee effort less monitoring induces 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**
[07] Irlenbusch and Ruchala (2008)	German Students (192)	Public Goods Game	• An external form of enforcement: Team-based compensation with and without a reward for the highest contributor in the team • The 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** (See text)
[08] Ariely, Bracha, and Meier (2009)	U.S. students (161)	Charity giving based on task performance	• An external form 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 effort is 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, van Dijk, De Cremer, et al. (2006).

Table 3. Bad news: Incentives provide information about the person who implements the incentive (I) (Continued…)

Citation	Subjects (number)	Games or activities	Institutional environments (treatments)	Results relevant to separability	Comment (quotes are from the cited paper)
[09] Stanca, Bruni, and Corazzini (2009)	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 to be in response to actions that are perceived as extrinsically motivated. **F**
[10] Fehr and Gächter (2002b) *	Swiss students (182)	Gift-Exchange Game	• Three external forms of enforcement: A Trust (pure fixed wage) contract, a 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 4. Moral disengagement: Incentives may suggest permissible behavior (F)

Citation	Subjects (number)	Games or activities	Institutional environments (treatments)	Results relevant to separability	Comment (quotes are from the cited paper)
[11] Hoffman, et al. (1994)	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-interested. **S**
[12] Schotter, et al. (1996)	U.S. students (247)	Ultimatum Game;	• 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 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-interested manner. "…the competition inherent in markets and the need to survive offers justifications for actions that, in isolation, would be unjustifiable" (p.38). **S**
[13] Cardenas, Stranlund, and Willis (2000)	Colombian forest area dwellers (112)	Common Pool Resource Game	• External enforcement device with a low-probability inspection and a fine • Communication	Fines induce more self-interested behavior and are ineffective at reducing common pool overexploitation in the longer run. Socially optimal individual 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 appears to have induced a shift from moral to self-interested frame. See Tenbrunsel and Messick (1999).
[14] Cardenas (2004)	Colombian users of rural ecosystems (265)	Common Pool Resource Game	• Different levels of external enforcement (low and high fines) 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 (compared to low) fines is less than 6 percent of the predicted effect assuming self-regarding preferences.	Marginal Crowding Out. (See text and also Table 7; where Categorical Crowding In also occurs).

Table 4. Moral disengagement: Incentives may suggest permissible behavior (F) (Continued…)

Citation	Subjects (number)	Games or activities	Institutional environments (treatments)	Results relevant to separability	Comment (quotes are from the cited paper)
[15] Heyman and Ariely (2004)	240 US students (150+90)	A computer task and a puzzle task	• Different forms of compensation (cash, candy or a cash-denominated amount of candy) • Different levels of monetary compensation (none, low, medium)	In both the cash and the candy conditions, effort increases when the compensation level increases from low to medium. In the no-compensation treatment, effort 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.	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.
[16] Bohnet and Baytelman (2007)	Senior executives in U.S. (353)	Trust Game and a Dictator Game	• No communication, face-to-face pre-play communication or post-play communication • An external form of enforcement (Post-play monetary punishment or not) • Stranger and Partner	Repetition and communication increase amount transferred and back-transferred; the option of punishment for low back-transfers reduces back-transfers of other-regarding trustees (those who send more in the Dictator Game).	"The availability of punishment destroys intrinsic trust and lowers people's willingness to reward trust" (p.1) **I**
[17] Houser, Xiao, McCabe, et al. (2008)	U.S. students (532)	Gift-Exchange Game	• A form of enforcement (Punishment as an incentive contract (i.e. a fine)) • Intention treatment: Punishment is assigned exogenously 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) and Mulder, et al. (2006) **I**
[18] Mellstrom and Johannesson (2008)	Swedish students (262)	Subjects are offered the opportunity to take a 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 prospective blood donors from 52% to 30% among women. No effect among men. Allowing individuals to donate the payment to charity eliminates 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. **S**

Table 4. Moral disengagement: Incentives may suggest permissible behavior (F) (Continued…)

Citation	Subjects (number)	Games or activities	Institutional environments (treatments)	Results relevant to separability	Comment (quotes are from the cited paper)
[19] Li, et al. (2009)	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 significantly increasing activities in parietal cortex previously implicated in economic 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**
[20] Henrich, et al. (2010) and Barr, et al. (2009)* and personal communication from Barr and Henrich (March 2009)	15 societies Including US students, African workers, Amazonian, Arctic, and African Hunter-gatherers. (428)	Dictator Game, Ultimatum Game and Third-Party Punishment Game (TPG)	• Differences between societies • Subjects played in the following sequence keeping their role (active or passive): first DG, then the UG and finally the TPG (an explicit incentive, i.e. fine)	In many populations in the TPG the incentives provided by the fine do not induce higher offers, but rather have the opposite effect; factors that may influence self-interest calculations (i.e. wealth, income and household size) are significant predictors of allocations in the TPG (but not in the DG). Membership in a 'world religion' positively associated with offers in the DG but not in the TPG	The presence of the fine in the TPG appears to have reduced the salience of moral reasoning (derived from the teachings of the world religions) and enhanced subjects concerns with their own economic needs. (See text)

Table 5. Control aversion: 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)
[21] Gneezy and Rustichini (2000b)	Israeli students (160 for the main experiment)	50 IQ test questions (plus a Principal Agent Game)	• 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) **F**
[22] Gneezy and Rustichini (2000b)	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 crowding out. See Gneezy (2003)
[23] Rustrom (2002)	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. Categorical crowding out.
[24] Falk and Kosfeld (2006)	Swiss students (804)	Gift-Exchange Game	• Different levels of a lower bound of performance selected by the experimenter (low, medium, and high) • The principal could choose whether to impose the minimum level or not • The principal chooses the agent's wage and whether to impose the bound	Most agents perform minimally (namely at the lower bound) in response to the principals' controlling decision. Majority of the principals anticipate this and do impose the bound, earning higher profits as a result.	Imposing a lower bound compromises subject's sense of autonomy and signals distrust and low expectations that diminish agents' reciprocity and good will towards the principal. Categorical crowding out. (See text) **I**
[25] Xiao and Houser (2011)	U.S. students (72)	Public Goods Game	• Exogenous punishment: None, private (only the punished subject knows when a round is monitored and the amount of the resulting punishment), public (all members of a group are told that information)	Private punishment induces lower levels of contribution than public punishment.	Weak incentives crowd out cooperation when implemented privately, but the same incentives when implemented publicly (but anonymously) promote cooperation.
[26] Gneezy (2003)*	US 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) and (2000b) and Mulder, et al. (2006). Categorical crowding out. **F**

Table 6. The economy produces people: Incentives alter how 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)
[27] Falkinger, et al. (2000) and personal communication from Gächter 18 February 2008.	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 not 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**
[28] Gneezy and Rustichini (2000a)	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 persists even after the fine is removed. No change in the four control centers.	The modest fine may have signaled 'how bad' lateness is and/or is perceived as a price of a service and displaces an ethical frame by a strategic one: "A fine is a price." **I, F, S**
[29] Bohnet, Frey, and Huck (2001)	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**
[30] Meier (2007)	Swiss students (11379)	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. But those who experience matching are substantially less likely to make a contribution to either fund in subsequent periods; 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. **F**

Table 6. The economy produces people: Incentives alter how new preferences are learned (E) (Continued…)

Citation	Subjects (number)	Games or activities	Institutional environments (treatments)	Results relevant to separability	Comment (quotes are from the cited paper)
[31] Reeson and Tisdell (2008)	Australian Students (98)	Public Goods Game	• Three external forms of enforcement: - 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 binding 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 occurs. 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.	Suasion enhances and imposed minimum contribution reduces other regarding preferences. Categorical crowding out. **F, S**
[32] Burks, et al. (2009)	Swiss (139) and US (113) bike messengers	Sequential Prisoners' Dilemma Game	Messenger exposure to performance based pay in their work place 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 learned under the incentive conditions of the work place are adopted outside the workplace.
[33] Irlenbusch and Sliwka (2005)*	German students (84)	Gift-Exchange Game	• Two internal forms of enforcement: The principal can choose - 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**
[34] Herrmann and Orzen (2008)*	British students (116)	Tullock Rent-Seeking Game or individual choice task and then a Prisoner's Dilemma Game	• Two different sequences (strategic vs. individual): • First stage: the two-player Tullock Rent-Seeking Game (with a different subject) or an individual choice task (with the same incentives). Second stage: a Prisoner's Dilemma	Players cooperate more when they previously played an individual choice task than when the previous game is competitive –strategic, one (i.e. the Rent-seeking Game) Cooperation and reciprocity rates decrease after subjects are exposed to rent-seeking competition.	Subjects may perceive the interaction in the rent-seeking contest as a negative one. "…an individual's attitude towards others undergoes changes between different types of situations because they evoke different contextual cues". (p. 3) "the experience of over-competitiveness in the contest game creates a disposition of rivalry in subjects that some cannot immediately "turn off" when the experiment ends" (p. 26)

Table 6. The economy produces people: Incentives alter how new preferences are learned (E) (Continued…)

Citation	Subjects (number)	Games or activities	Institutional environments (treatments)	Results relevant to separability	Comment (quotes are from the cited paper)
[35] Gächter, Kessler, and Konigstein (2010)*	Swiss students (500)	Gift-Exchange Game	• Three external forms of enforcement: a Trust (pure fixed wage contract), a 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) i.e. there is no voluntary cooperation. Experiencing well-designed contracts reduces voluntary cooperation even after incentives are withdrawn.	Incentives may have a lasting negative effect on voluntary cooperation. **F**

Table 7. Incentives crowd in social preferences (C)

Citation	Subjects (number)	Games or activities	Institutional environment (treatments)	Results relevant to separability	Comment (quotes are from the cited paper)
[36] Falk, Gächter, and Kovacs (1999)	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).
[37] Gächter and Falk (2002)	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 then 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)
[38] Masclet, et al. (2003)	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 monetary and non-monetary sanctions induce 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 Punishment device is removed, having previous monetary sanctions show higher contributions than having non-monetary sanctions.	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 payoff maximizing subject, nonetheless 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, Murphy, Spraggon, et al. (2011)
[-]* Cardenas (2004)	Colombian users of rural ecosystems (265)	Common Pool Resource Game	• Different levels of external enforcement (low and high fines) with announcement of socially optimal extraction level and without communication • Communication without fines and announcement.	Deviation from the self-regarding Nash extraction level was 29% greater under the small fine than with no fine.	Individuals consider the norm of cooperation that is proposed externally [the announced optimal level] when extracting (p. 238). Categorical Crowding In. (See text and also Table 3; where Marginal Crowding out also occurs)

* This reference is not numbered since it is an additional result of an already cited study [14].

Table 7. Incentives crowd in social preferences (C) (Continued…)

Citation	Subjects (number)	Games or activities	Institutional environment (treatments)	Results relevant to separability	Comment (quotes are from the cited paper)
[39] Henrich, et al. (2005)	Foragers, herders, others in 15 small-scale societies (1128)	Ultimatum Game	• 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). This study also presents evidence of incentives alter how new preferences are learned **E**
[40] 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 (2008), Vertova and Galbiati (2010)
[41] Tyran and Feld (2006)	Swiss students (102)	Public Goods Game	• Levels of sanctions: none, mild and severe • Enforcement: external (i.e. experimenter-imposed) or self-imposed (by referendum)	Experimenter imposed mild sanctions do not significantly affect average contributions to the public good. Compliance is much improved if mild law is endogenously chosen.	Experimenter imposed sanctions raised the expected cost of freeriding without affecting behavior; only referendum imposed sanctions conveyed a signal of moral disapproval by peers.
[42] Herrmann, et al. (2008a)	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 (targeting 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 so that peer punishment induces shame rather than resentment. The quality of the formal law enforcement institutions and informal sanctions are complements, "because antisocial punishment is lower in these societies." (p. 1367).
[43] Rodriguez-Sickert, Guzmán, and Cárdenas (2008)	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 by withdrawing cooperation.

Table 7. Incentives crowd in social preferences (C) (Continued…)

Citation	Subjects (number)	Games or activities	Institutional environment (treatments)	Results relevant to separability	Comment (quotes are from the cited paper)
[44] Carpenter, et al. (2009)	US students (172)	Public Goods Game	• Costly punishment: subjects can punish non-cooperators at a cost to themselves • Different team's residual claim (marginal per capita return on the public good) • Different group size	Shirkers are punished by peers and they 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 peer-imposed incentives. This study also presents evidence of incentives alter how new preferences are learned **E**
[45] Carpenter and Myers (2010)	U.S. Volunteer firefighters (217) and non-volunteer community members (189)	Dictator Game	• Exogenous variation in the presence and level of small stipends paid to volunteer firefighters	Small monetary incentives increase turnout to fighting fires for firemen unconcerned about image but have no effect on image-concerned firemen (the estimated negative effect is not significant).	The effect of image concerns increases with the visibility of the activity (training is a less visible activity than fighting fires). For firefighters with image concerns the positive direct effect of small extrinsic incentives is canceled by the negative indirect effect of incentives on the image-value of fighting fires.
[46] Gächter, Nosenzo, and Sefton (2010)	British Students (84)	Gift-Exchange Game with 3-member firms (one employer and two employees)	• Employees move sequentially (Employee 1 has pay comparison information (i.e. information about what coworker earns) and Employee 2 additionally has an effort comparison information (information about how co-worker performs) • Employers can 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 co-worker that provides less effort. Reciprocity toward the employer 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 work effort and hence deservingness, not to employer generosity.
[47] Lopez, et al. (2011)	Colombian Fishermen (180)	Public Goods Game	• Public reminder about benefits of cooperation plus 1/5 chance of receiving private reminder of the social losses resulting from the individual's non-cooperative behavior (Guilt); receiving public reminder of the social losses resulting from the individual's non-cooperative behavior (Shame), facing an external low penalty for not contributing to the public good, facing an external high penalty for not contributing.	Priming subjects to feel guilty about low contributions did not affect average contributions, but the random public revelation of one's contributions (inducing shame) substantially increased contributions. Experimenter's imposition of the fine further increased contributions but the level of the fine has no effect.	Results suggest the importance of moral framing and that the fine did not work as an incentive but rather as a signal highlighting the salience of the ethical dimension of the problem. Categorical Crowding in.

Table 7. Incentives crowd in social preferences (C) (Continued…)

Citation	Subjects (number)	Games or activities	Institutional environment (treatments)	Results relevant to separability	Comment (quotes are from the cited paper)
[48] Vertova and Galbiati (2010)*	Italian students (210)	Public Goods Game (and a Lottery Game)	• Different levels of a stated (non-binding) obligation to contribute (zero, low and high) • A symmetric incentive structure (a level of contribution less (more) than the minimum contribution could be subject to a probabilistic penalty (reward)) with low and medium size	There is a positive effect of the obligation which is greater when it is combined with a weak monetary incentive than when no incentives are offered. A stronger monetary incentive does not result in an increase in contributions. The strong monetary incentive also has no effect on behavior in the absence of the stated obligation.	Incentives not only influence material payoffs but also frame recommended high contributions as obligations. Incentives and obligations tie up people's behaviors by activating values and/or coordinating individuals' beliefs. See also Galbiati and Vertova (2008)
[49] Barr (2001)*	Zimbabwean villagers (602)	Public Goods Game	• Two external forms of non-monetary punishment i)Public announcement: each player announces her level of contribution to everyone present in the session, ii) 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 maximizing were unaffected). See Gächter 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.
[50] 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 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 official's decision to engage in bribery." (p.17)

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.

Akerlof, G. A. and R. Kranton. 2010. *Identity Economics: How our identities shape our work, wages, and well-being*. Princeton: Princeton University Press.

Al-Ubaydli, O., D. Houser, J. Nye, M. P. Paganelli, and X. S. Pan. 2011. "The causal effect of market participation on trust: An experimental investigation using randomized control " *Mimeo*.

Andreoni, J. 1990. "Impure Altruism and Donations to Public Goods: A Theory of Warm-Glow Giving." *THe Economic Journal*, 100:401, pp. 464-77.

Angrist, J. and V. Lavy. 2009. "The effects of high stakes high school achievement rewards: Evidence from a randomized trial." *American Economic Review*, 99:4, pp. 1384-414.

Ariely, D., A. Bracha, and S. Meier. 2009. "Doing Good or Doing Well? Image Motivation and Monetary Incentives in Behaving Prosocially." *American Economic Review*, 99:1, pp. 544-55.

Ariely, D., U. Gneezy, G. Loewenstein, and N. Mazar. 2009. "Large Stakes and Big Mistakes." *Review of Economic Studies*, 76:2, pp. 451-69.

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

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

Bandiera, O., I. Barankay, and I. Rasul. 2005. "Social Preferences and the Response to Incentives: Evidence from Personnel Data." *The Quarterly Journal of Economics*, 120:3, pp. 917-62.

Bandura, A. 1991. "Social congnitive theory of moral thought and action," in *Handbook of Moral Behavior and Development: Volume I, Theory*. Kurtines and Gewirtz eds. Hillsdale, New Jersey: Lawrence Erlbaum and Associates, pp. 45-103.

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

Baran, N. M., P. Sapienza, and L. Zingales. 2010. "Can we infer social preferences from the lab? Evidence from the trust game." *NBER Working Papers*.

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. 2004. "The effects of birthplace and current context on other-regarding preferences." *University of Oxford, Centre for the Study of Africa Economies*.

Barr, A., C. Wallace, J. Ensminger, J. Henrich, C. Barrett, A. Bolyanatz, J. C. Cardenas, M. Gurven, E. Gwako, C. Lesorogol, F. Marlowe, R. McElreath, D. Tracer, and J. Ziker. 2009. "Homo Æqualis: A Cross-Society Experimental Analysis of Three Bargaining Games." *Discussion Paper Series*. Department of Economics, University of Oxford: Oxford.

Becker, G. S. 1976. "Altruism, Egoism, and Genetic Fitness: Economics and Sociobiology." *Journal of Economic Literature*, 14:3, pp. 817-26.

Belkin, D. 2002. "Boston Firefighters Sick - Or Tired of Working." *Boston Globe*, 18 January, Third ed.: B1: Boston.

Bellemare, C., S. Kröger, and A. Van Soest. 2008. "Measuring Inequity Aversion in a Heterogeneous Population Using Experimental Decisions and Subjective Probabilities." *Econometrica*, 76:4, pp. 815-39.

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. 1970 [1789]. *An Introduction to the Principles of Morals and Legislation*: Athlone Press.

Benz, M. and S. Meier. 2008. "Do people behave in experiments as in the field?—evidence from donations." *Experimental Economics*, 11:3, pp. 268-81.

Besley, T., G. Bevan, and K. Burchardi. 2009. "Naming & Shaming: The impacts of different regimes on hospital waiting times in England and Wales'." *CEPR Discussion Paper no. 7306.* Centre for Economic Policy Research. : London.

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

Birch, L. L. and D. W. Marlin. 1982. "I don't like it; I never tried it: effects of exposure on two-year-old children's food preferences." *Appetite*, 3:4, pp. 353-60.

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.

Bisin, A. and T. Verdier. 2011. "The Economics of Cultural Transmission and Socialization," in *Handbook of Social Economics*. Benhabib, Bisin and Jackson eds. The Netherlands: North-Holland Elsevier, pp. 339-416.

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.

Bloom, P. 2010. *How pleasure works: The new science of why we like what we like*. New York: Norton.

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.

Bohnet, I., F. Greig, B. Herrmann, and R. Zeckhauser. 2008. "Betrayal Aversion: Evidence from Brazil, China, Oman, Switzerland, Turkey, and the United States." *American Economic Review*, 98:1, pp. 294-310.

Bohnet, I., B. Herrmann, M. Al-Ississ, A. Robbett, K. Al-Yahia, and R. Zeckhauser. 2010. "The Elasticity of Trust: How to Promote Trust in the Arab Middle East and the United States." *Faculty Research Working Paper Series*. Harvard Kennedy School: Cambridge.

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. 2011. "Is Liberal Society a Parasite on Tradition?" *Philosophy & Public Affairs*, 39:1, pp. 46-81.

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

Bowles, S. and H. Gintis. 2011. *A cooperative species: human reciprocity and its evolution*. Princeton: Princeton University Press.

Bowles, S. and S.-H. Hwang. 2008. "Social preferences and public economics: Mechanism design when social preferences depend on incentives." *Journal of Public Economics*, 92:8-9, pp. 1811-20.

Boyd, R. and P. J. Richerson. 1985. *Culture and the Evolutionary Process*. Chicago: University of Chicago Press.

Burks, S., J. Carpenter, and L. Goette. 2009. "Performance pay and worker cooperation: Evidence from an artefactual field experiment." *Journal of Economic Behavior & Organization*, 70:3, pp. 458-69.

Cardenas, J. 2011. "Social Norms and Behavior in the Local Commons as Seen Through the Lens of Field Experiments." *Environmental and Resource Economics*, pp. 1-35.

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. 2005. "Groups, Commons and Regulations: Experiments with Villagers and Students in Colombia," in *Psychology, Rationality and Economic Behaviour: Challenging Standard Assumptions*. Agarwal and Vercelli eds: International Economics Association.

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. 2009. "Strong Reciprocity and Team Production: Theory and Evidence." *Journal of Economic Behavior and Organization*, In press.

Carpenter, J., C. Connolly, and C. Myers. 2008. "Altruistic behavior in a representative dictator experiment." *Experimental Economics*, 11:3, pp. 282-98.

Carpenter, J. and C. K. Myers. 2010. "Why volunteer? Evidence on the role of altruism, image, and incentives." *Journal of Public Economics*, 94:11-12, pp. 911-20.

Carpenter, J., E. Verhoogen, and S. Burks. 2005. "Comparing students to workers: The effects of social framing on behavior in distribution games." *Research in Experimental Economics Series*, Vol. 10: 393-98.

Carpenter, J. P. and E. Seki. 2010. "Do social preferences increase productivity? Field experimental evidence from fishermen in Toyama Bay." *Economic Inquiry*, in press.

Cavalli-Sforza, L. L. and M. W. Feldman. 1981. *Cultural transmission and evolution : a quantitative approach*. Princeton, N.J.: Princeton University Press.

Cervellati, M., J. Esteban, and L. Kranich. 2010. "Work values, endogenous sentiments redistribution." *Journal of Public Economics*, 94:9-10, pp. 612-27.

Cleave, B. L., N. Nikiforakis, and R. Slonim. 2010. "Is There Selection Bias in Laboratory Experiments?" *Economics Working Paper Series*. University of Sidney: Sidney.

Cohn, A., E. Fehr, and L. Goette. 2011. "Fair Wages and Effort: Evidence from a Field Experiment." *Mimeo*.

Cooley, C. H. 1902. *Human nature and social order*. New York: Scribner's.

Cooter, R. 2000. "Do Good Laws Make Good Citizens? An Economic Analysis of Internalized Norms." *Virginia Law Review*, 86:8, pp. 1577-601.

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

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.

DellaVigna, S. 2009. "Psychology and Economics: Evidence from the Field." *Journal of Economic Literature*, 47:2, pp. 315-72.

Dickenson, D. and M.-C. Villeval. 2008. "Does monitoring decrease work effort? The complementarity between agency and crowding-out theories." *Games and Economic Behavior*, 63:1, pp. 56-76.

Drago, F., R. Galbiati, and P. Vertova. 2009. "The Deterrent Effects of Prison: Evidence from a Natural Experiment." *Journal of Political Economy*, 117:2, pp. 257-80.

Dufwenberg, M., P. Heidhues, G. Kirchsteiger, F. Riedel, and J. Sobel. 2011. "Other-regarding preferences in general equilibrium " *Review of Economic Studies*, (Published online: February 4, 2011).

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, J. Möllerström, and S. Munkhammar. 2011. "Social Framing Effects: Preferences or Beliefs? ." Stockholm School of Economics, Ramböll Management and Harvard University. At: http://www2.hhs.se/personal/ellingsen/pdf/FramingGEB1.pdf.

Elster, J. 1998. "Emotions and Economic Theory." *Journal of Economic Literature*, 36:1, pp. 47-74.

Ertan, A., T. Page, and L. Putterman. 2009. "Who to punish? Individual decisions and majority rule in mitigating the free rider problem." *European Economic Review*, 53:5, pp. 495-511.

Falk, A., E. Fehr, and U. Fischbacher. 2005. "Driving Forces behind Informal Sanctions." *Econometrica*, 73:6, pp. 2017-30.

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. Gächter, 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 J. J. Heckman. 2009. "Lab Experiments Are a Major Source of Knowledge in the Social Sciences." *Science*, 326:5952, pp. 535-38.

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

Falk, A., S. Meier, and C. Zehnder. 2011. "Did We Overestimate the Role of Social Preferences? The Case of Self-Selected Student Samples." *IZA Discussion Papers*.

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

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

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

Fehr, E., S. Gächter, and G. Kirchsteiger. 1997. "Reciprocity as a Contract Enforcement Device: Experimental Evidence." *Econometrica*, 65:4, pp. 833-60.

Fehr, E. and L. Goette. 2007. "Do Workers Work More if Wages Are High? Evidence from a Randomized Field Experiment." *American Economic Review*, 97:1, pp. 298.

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

Fehr, E. and A. Leibbrandt. 2011. "A field study on cooperativeness and impatience in the Tragedy of the Commons." *Journal of Public Economics*, In Press, Corrected Proof.

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. 1999. "A Theory of Fairness, Competition, and Cooperation." *Quarterly Journal of Economics*, 114:3, pp. 817-68.

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.

Fiske, A. P. 1992. "The Four Elementary Forms of Sociality: Framework for a Unified Theory of Social Relations." *Psychological Review*, 99:4, pp. 689-723.

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. 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.

Frey, B. S. and F. Oberholzer-Gee. 1997. "The Cost of Price Incentives: An Empirical Analysis of Motivation Crowding-Out." *American Economic Review*, 87:4, pp. 746-55.

Frey, B. S. and A. Stutzer. 2008. "Environmental morale and motivation," in *The Cambridge Handbook of Psychology and Economic Behaviour*. Lewis ed. Cambridge: Cambridge University Press.

Fryer, R. 2011. "Financial incentives and student acheivement: Evidence from randomized trials." *Quarterly Journal of Economics*, 126:4.

Funk, P. 2007. "Is There An Expressive Function of Law? An Empirical Analysis of Voting Laws with Symbolic Fines." *American Law and Economics Review*.

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

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

Gächter, S., E. Kessler, and M. Konigstein. 2010. "Do incentives destroy Voluntary Cooperation?" *University of Nottingham, School of Economics*: Nottingham.

Gächter, S., D. Nosenzo, and M. Sefton. 2010. "The Impact of Social Comparisons on Reciprocity." *Discussion Papers*, Vol. 2010-10. The Centre for Decision Research and Experimental Economics, School of Economics, University of Nottingham. Forthcoming in Scandinavian Journal of Economics.

Galbiati, R. and P. Vertova. 2008. "Obligations and cooperative behaviour in public good games." *Games and Economic Behavior*, 64:1, pp. 146-70.

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.

Grant, R. 2011. *Strings attached: Untangling the ethics of incentives.* . Princeton.

Greenberger, S. 2003. "Sick day abuses focus of fire talks." *Boston Globe*, 17 September, Third ed.: B7.

Guth, W. and H. Kliemt. 1994. "Competition or Co-operation: On the Evolutionary Economics of Trust, Exploitation, and Moral Attitudes." *Metroeconomica*, 45:2, pp. 155-87.

Healy, K. 2006. *Best Last Gifts: Altruism and the Market for Human Blood and Organs*. Chicago: University of Chicago Press.

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

Helsley, R. W. and W. C. Strange. 2000. "Social Interactions and the Institutions of Local Government." *The American Economic Review*, 90:5, pp. 1477-90.

Henrich, J., R. Boyd, S. Bowles, C. Camerer, E. Fehr, H. Gintis, R. McElreath, M. Alvard, A. Barr, J. Ensminger, N. S. Henrich, K. Hill, F. Gil-White, M. Gurven, F. Marlowe, J. Patton, and D. Tracer. 2005. "'Economic Man' in Cross-Cultural Perspective: Behavioral experiments in 15 small-scale societies." *Behavioral and Brain Sciences*, 28, pp. 795-855.

Henrich, J., J. Ensminger, R. McElreath, A. Barr, C. Barrett, A. Bolyanatz, J. C. Cardenas, M. Gurven, E. Gwako, N. Henrich, C. Lesorogol, F. Marlowe, D. Tracer, and J. Ziker. 2010. "Markets, Religion, Community Size, and the Evolution of Fairness and Punishment." *Science*, 327:5972, pp. 1480-84.

Herrmann, B. and H. Orzen. 2008. "The appearance of homo rivalis: Social preferences and the nature of rent seeking." *Discussion Papers 2008-10*. The Centre for Decision Research and Experimental Economics, School of Economics, University of Nottingham: Nottingham.

Herrmann, B., C. Thoni, and S. Gächter. 2008a. "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.

Hirschman, A. O. 1985. "Against parsimony:three ways of complicating some categories of economic discourse." *Economics and Philosophy*, 1:1, pp. 7-21.

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.

Holmås, T. H., E. Kjerstad, H. Lurås, and O. R. Straume. 2010. "Does monetary punishment crowd out pro-social motivation? A natural experiment on hospital length of stay." *Journal of Economic Behavior & Organization*, 75:2, pp. 261-67.

Hopfensitz, A. and E. Reuben. 2009. "The Importance of Emotions for the Effectiveness of Social Punishment*." *The Economic Journal*, 119:540, pp. 1534-59.

Houser, D., E. Xiao, K. McCabe, and V. Smith. 2008. "When Punishment Fails: Research on Sanctions, Intentions, and Non-Cooperation." *Games and Economic Behavior*, 62:2, pp. 509-32.

Hwang, S.-H. and S. Bowles. 2011a. "Mechanism design with state-dependent preferences." *Mimeo.*: Santa Fe Institute.

Hwang, S.-H. and S. Bowles. 2011b. "The sophisticated planner's dilemma: optimal incentives with endogenous preferences." *Santa Fe Institute*: Santa Fe.

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* http://ssn.com/abstract=822866.

Karlan, D. 2005. "Using Experimental Economics to Measure Social Capital and Predict Financial Decisions." *American Economic Review*, 95:5, pp. 1688-99.

Kessler, E. 2008. "Behavioural Economics of Performance Incentives." *School of Economics*, Vol. PhD Economics. University of Nottingham: Nottingham.

Kosfeld, M., A. Okada, and A. Riedl. 2009. "Institution Formation in Public Goods Games." *American Economic Review*, 99:4, pp. 1335-55.

Kreps, D. M. 1997. "Intrinsic Motivation and Extrinsic Incentives." *The American Economic Review*, 87:2, pp. 359-64.

Kroll, S., T. L. Cherry, and J. F. Shogren. 2007. "Voting, punishment and public goods." *Economic Inquiry*, 45:3, pp. 557-70.

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.

Lazear, E. 2000. "Performance Pay and Productivity." *American Economic Review*, 90:5, pp. 1346 - 61.

Leibbrandt, A., U. Gneezy, and J. List. 2010. "Ode to the Sea: The socio-ecological underpinnings of social norms." *University of Chicago, Department of Economics*.

Leider, S., M. M. Möbius, T. Rosenblat, and Q.-A. Do. 2009. "Directed Altruism and Enforced Reciprocity in Social Networks." *The Quarterly Journal of Economics*, 124:4, pp. 1815-51.

Lepper, M. R., G. Sagotsky, J. L. Dafoe, and D. Greene. 1982. "Consequences of superfluous social constraints: Effects on young children's social inferences and subsequent intrinsic interest." *Journal of Personality and Social Psychology*, 42:1, pp. 51-65.

Leung, K. T. and J. L. Martin. 2003. "The Looking Glass Self: An Empirical Test and Elaboration." *Social Forces*, 81:3, pp. 843-79.

Levine, D. K. 1998. "Modeling Altruism and Spitefulness in Experiments." *Review of Economic Dynamics*, 1:3, pp. 593-622.

Levitt, S. D. and J. List. 2007. "What do laboratory experiments measuring social preferences reveal about the real word." *Journal of Economic Perspectives*, 21:1, pp. 153-74.

Li, J., E. Xiao, D. Houser, and P. R. Montague. 2009. "Neural responses to sanction threats in two-party exchanges." *Proceedings of the National Academy of Sciences U S A*, 106:39, pp. 16835-40.

List, J. 2004. "Young, Selfish, and Male: Field evidence on social preferences." *Economic Journal*, 114, pp. 121-49.

Loewenstein, G. 2000. "Emotions in Economic Theory and Economic Behavior." *American Economic Review*, 90:2, pp. 426-32.

Lopez, M. C., J. J. Murphy, J. M. Spraggon, and J. K. Stranlund. 2011. "Comparing the Effectiveness of Regulation and Prosocial Emotions to Enhance Cooperation: Experimental Evidence from Fishing Communities in Colombia." *Economic Inquiry*, doi: 10.1111/j.1465-7295.2010.00344.x, pp. no-no.

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

MacLeod, W. B. 2007. "Can Contract Theory Explain Social Preferences?" *American Economic Review*, 97:2, pp. 187-92.

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.

Mellizo, P., J. Carpenter, and P. H. Matthews. 2011. "Workplace Democracy in the Lab." *Mimeo*.

Mellstrom, C. and M. Johannesson. 2008. "Crowding Out in Blood Donation: Was Titmuss Right?" *Journal of The European Economic Association*, 6:4, pp. 845-63.

Mill, J. S. 1867[1848]. *Principles of Political Economy with Some of Their Applications*. London: Longmass, Green, Reader, and Diver.

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.

Murphy, S. T., J. L. Monahan, and R. B. Zajonc. 1995. "Additivity of Nonconscious Affect: Combined Effects of Priming and Exposure." *Journal of Personality and Social Psychology*, 69:4, pp. 589-602.

Murphy, S. T. and R. B. Zajonc. 1993. "Affect, Cognition, and Awareness: Affective Priming With Optimal and Suboptimal Stimulus Exposures." *Journal of Personality and Social Psychology*, 64:5, pp. 723-39.

Rabin, M. 1993. "Incorporating Fairness into Game Theory and Economics." *American Economic Review*, 83:5, pp. 1281-302.

Rabin, M. 1998. "Psychology and Economics." *Journal of Economic Literature*, 36:1, pp. 11-46.

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.

Rigdon, M. 2009. "Trust and reciprocity in incentive contracting." *Journal of Economic Behavior & Organization*, 70:1-2, pp. 93-105.

Rodriguez-Sickert, C., R. A. Guzmán, and J. C. Cárdenas. 2008. "Institutions influence preferences: Evidence from a common pool resource experiment." *Journal of Economic Behavior & Organization*, 67:1, pp. 215-27.

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.

Rustagi, D., S. Engel, and M. Kosfeld. 2010. "Conditional Cooperation and Costly Monitoring Explain Success in Forest Commons Management." *Science*, 330:6006, pp. 961-65.

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

Schnedler, W. and R. Vadovic. 2011. "Legitimacy of Control." *Journal of Economics and Management Strategy*, Forthcoming.

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.

Seabright, P. 2009. "Continuous Preferences and Discontinous Choices: How Altruists Respond to Incentives." *The B.E. Journal of Theoretical Economics*, 9:Article 14.

Serra, D. 2008. "Combining Top-down and Bottom-up Accountability: Evidence from a Bribery Experiment." *Centre for the Study of African Economies Series*. University of Oxford. : Oxford.

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

Shu, L., F. Gino, and M. H. Bazerman. 2009. "Dishonest Deed, Clear Conscience: Self-preservation through moral disengagement and motivated forgetting."

Sliwka, D. 2007. "Trust as a Signal of a Social Norm and the Hidden Costs of Incentive Schemes." *American Economic Review*, 97:3, pp. 999-1012.

Sloof, R. and J. Sonnemans. 2011. "The interaction between explicit and relational incentives: An experiment." *Games and Economic Behavior*, In Press, Accepted Manuscript.

Sobel, J. 2002. "Can We Trust Social Capital?" *Journal of Economic Literature*, 40:1, pp. 139-54.

Sobel, J. 2005. "Interdependent Preferences and Reciprocity." *Journal of Economic Literature*, 43:2, pp. 392-436.

Sobel, J. 2009. "Generous actors, selfish actions: markets with other-regarding preferences." *International Review of Economics*, 56:1, pp. 3-16.

Sobel, J. 2010. "Markets and Other-Regarding Preferences." Department of Economics, University of California. : San Diego. At: http://federation.ens.fr/ydepot/semin/texte0910/SOB2010MAR.pdf.

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. 2009. "Testing theories of reciprocity: Do motivations matter?" *Journal of Economic Behavior & Organization*, 71:2, pp. 233-45.

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

Sunstein, C. R. 1996. "On the Expressive Function of Law." *University of Pennsylvania Law Review*, 144:5, pp. 2021-53.

Sutter, M., S. Haigner, and M. G. Kocher. 2011. "Choosing the Carrot or the Stick? Endogenous Institutional Choice in Social Dilemma Situations." *Review of Economic Studies*, 77:4, pp. 1540-66.

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.

Titmuss, R. M. 1971. *The Gift Relationship: From Human Blood to Social Policy*. New York: Pantheon Books.

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.

Vertova, P. and R. Galbiati. 2010. "How Laws Affect Behavior: Obligations, Incentives and Cooperative Behavior." *SSRN eLibrary*: At: http://ssrn.com/paper=1615349.

Warneken, F. and M. Tomasello. 2008. "Extrinsic rewards undermine altruistic tendencies in 20-month-olds." *Developmental Psychology*, 44:6, pp. 1785-88.

Xiao, E. and D. Houser. 2011. "Punish in public." *Journal of Public Economics*, 95:7-8, pp. 1006-17.

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., F. Gino, and V. Bohns. 2010. "Good lamps are the best police: Darkness increases dishonesty and self-interested behavior." *Pscyhological Science*, 21:3, pp. 311-14.

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.