Social Preferences and Public Economics: Are Good Laws a Substitute for Good Citizens?

Samuel Bowles

SFI WORKING PAPER: 2007-01-003

SANTA FE INSTITUTE

Social Preferences and Public Economics: Are good laws a substitute for good citizens?

Samuel Bowles
Santa Fe Institute and University of Siena[1]
14 January, 2007

Abstract

Laws and policies designed to harness self-regarding preferences to public ends may fail when they compromise the beneficial effects of pro-social preferences. Experimental evidence indicates that incentives that appeal to self interest may reduce the salience of intrinsic motivation, reciprocity, and other civic motives. Motivational crowding in also occurs. The evidence for these processes is reviewed and a model of optimal explicit incentives is presented.

JEL: D64, D52, H41, H21, Z13, C92

Keywords: Social preferences, implementation theory, incentive contracts, incomplete contracts, framing, behavioral experiments, motivational crowding out, ethical norms, constitutions

---

[1] Thanks to the Behavioral Sciences Program of the Santa Fe Institute and the University of Siena for financial support of this project, and the Certosa di Pontignano (Siena) for providing an ideal research environment. I would like to thank Margaret Alexander, Abigail Barr, Iris Bohnet, James Boyce, Juan Camilo Cardenas, Josh Cohen, Gerald Cohen, Kenan Ercel, Ernst Fehr, Simon Gaechter, John Geanakoplos, Sung Ha Hwang, Amara Moore-Levy, Suresh Naidu, Elinor Ostrom, Michael Kosfeld, John Roemer, Bob Rowthorn, Paul Seabright, Rajiv Sethi, Joaquim Silvestre, E. Somanathan, Tim Taylor, Elisabeth Wood, Erik Olin Wright and Giulio Zanella for their contributions to this research.

1. Introduction

Policies designed to harness self-regarding preferences to public ends may be counter-productive.[2] These failures occur when conventional self-interest-based policies compromise the beneficial effects of intrinsic motivation and reciprocity as well as civic virtues such as a concern for fairness and a desire to uphold social norms.

A recent quasi-natural experiment (Gneezy and Rustichini (2000a)) provides an example. In Haifa, at six randomly chosen day care centers, a fine was imposed on parents who were late in picking up their children at the end of the day (in a control group of centers no fine was imposed). Parents responded to the fine by significantly greater tardiness: the fraction late more than doubled. When after 16 weeks the fine was revoked, their enhanced tardiness persisted, showing no tendency to return to the *status quo ante*. Over the entire 20 weeks of the experiment there were no changes in the degree of lateness in the control group. While other interpretations are possible, the counter-productive imposition of the fines appears to illustrate a kind of crowding out: using a market mechanism (the fine) seems to have undermined the parents' sense of ethical obligation to avoid inconveniencing the teachers (Gneezy (2003)).

Other interventions may crowd *in* civic-mindedness. In the tradition of the *charivaris* of early modern Europe (Tilly (1981)), the municipal commissioner of the Indian city of Rajahmundry hired ten drummers and directed them to beat non-stop outside the homes and offices of tax evaders (Farooq (2005)). The policy was highly effective, apparently by inducing shame among the tax evaders by the public denunciation of their transgression of a social norm.

These examples point to a shortcoming in the conventional approach to policy implementation and constitutional design, reviewed in the next section. What appear to be improved incentives in this framework may have counter-productive effects due to motivational crowding out. (I use 'incentives' without adjective to mean incentives appealing to self-regarding preferences.) Crowding in may also occur, when incentives not only activate self- regarding motives to engage in socially valued behaviors, but also recruit other-regarding preferences to the same end. In section 3 I present experiment evidence showing that the effects of explicit incentives often depend on whether they enhance or diminish the salience

[2] Related views have been advanced by Titmuss (1971), Taylor (1976), Ben-Porath (1980), Hirschman (1985), Bowles (1989), Frohlich and Oppenheimer (1995), Kreps (1997) Frey (1997) Cooter (1998), Ostrom (2000), Seabright (2004), Benabou and Tirole (2005) and others.

of other-regarding preferences or intrinsic motivation. Optimal incentives in the presence of motivational crowding are defined in section 4. The concluding section proposes a revision of the standard approach to implementation.

2. Civic virtue and public policy

In his *Essays*: *Moral, Political and Literary* (1742) David Hume (1964):117-118 recommended that

> in contriving any system of government ... every man ought to be supposed to be a *knave* and to have no other end, in all his actions, than private interest. By this interest we must govern him, and, by means of it, make him, notwithstanding his insatiable avarice and ambition, cooperate to public good.

Hume's maxim that public policies should harness self-regarding preferences to public ends remains a foundation of public economics, its wisdom buttressed by ample evidence that conventional incentive-based contracts and policies often work very well (Laffont and Matoussi (1995), Lazear (2000)). The empirical importance of civic (other-regarding) motives for public economics has also long been recognized and has recently been suggested in studies of tax compliance (Pommerehne and Weck-Hannermann (1996), Andreoni, Erand, and Feinstein (1998)), political opinion and voting concerning income security and redistribution measures (Fong, Bowles, and Gintis (2005)), and generalized obedience to law (Kahan (1997)).

Hume, Bentham and the other classicals advocating self interest as a basis of public policy design did not ignore moral behavior, but instead assumed it would be unaffected by incentive-based policies designed to harness self-interest. Along with civic virtue, explicit incentives and constraints could thus contribute additively to good government. As a result of this implicit 'separability assumption' they failed to take account of the conditions under which civic virtue would flourish and favorably affect aggregate outcomes and how harnessing self interest to the public good might either attenuate or enhance civic virtue. Modern public economics, implementation theory, mechanism design and related fields continue this practice.

To clarify the notion of separability, consider a community of identical individuals who may contribute to a public goods project by taking an action ($a \in [0,1]$) at a cost $a^2/2$. The output of the project varies with the sum of the contributions of the members, and is divided such that each member receives an amount $\phi\sum a_k$ for $k = 1...n$ The public good technology is such that $\phi < 1 < n\phi$ so that in the absence of taxes or values the individual will contribute less than 1, the social optimum. Explicit incentives take the form of a tax at rate $t \in [0,1]$ on the shortfall of one's action from the social optimum (1- a).

I explore the various dimensions other-regarding preferences shortly, but for now I simply refer to these influences on behavior as 'values' and represent them by $v(a,t)$. Each individual's utility function is (suppressing the subscripts indexing the individual) is $u = \pi + v(a,t)$) respectively capturing self-regarding preferences over net payoffs to the public project (π) and values. Then

(1) $\quad u = - a^2/2 + \phi \sum a_k - t(1\text{-a}) + v(a,t)$

This determines the individual's best response a^* given by

(2) $\quad a^* = \phi + t + v_a(a, t)$

where the left hand side is the marginal cost of contributing and the remaining three (right hand side) terms are marginal costs arising from the private material benefits (from the project and from reduced taxes) and the value benefits.

The classical separability assumption maintains that the level of explicit material incentives does not influence the effectiveness of values in sustaining contributions: that is $\partial v_a / \partial t \equiv \delta = 0$. Where this separability condition does not hold, we have either crowding in ($\delta > 0$) where greater explicit incentives enhance the effects of values or crowding out ($\delta < 0$) where the opposite occurs. (The former is a type of super-modularity, while the latter represents sub-modularity).

When separability fails, this may be the result of one of the following causes.

Framing. Incentives may signal appropriate behavior shifting the frame from ethical and other-regarding to instrumental and self-regarding, or because the incentives provide a signal of the cost (to another) of the individual's behavior, in which case self-regarding behavior modified by the incentive would seem appropriate behavior. (Kahneman and Tversky (1986))

Information about intent or type. As Seabright (2004), Benabou and Tirole (2005) and Sliwka (2007) point out, any incentive selected by a principal inevitably conveys information about the principal's preferences or beliefs concerning the agent or the distribution of types among agents. Explicit incentives may provide a negative signal about the principal's type or beliefs, either in the form of lack of concern about the agent's well being or lack of trust.

Self-determination. Where intrinsic motivation is present, incentives may 'overjustify' the activity and reduce the individual's sense of autonomy. 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).)

Endogenous preferences.[3] Incentives may alter the duration, information structure, degree of positive assortment, face-to-face-ness and other aspects of social interactions. By altering who meets who to do what with what rewards, incentives affect the process by which individuals update their preferences, possibly leading to a long term shift in equilibrium endogenous preferences (Bowles (1998)).

3. Explicit incentives vs civic motives?

If only self-regarding motives are at work, the separability assumption cannot fail. The reason is that the policy maker is then working with a *tabla rasa*: the mobilization of self-regarding motives towards some public end cannot extinguish other motives that might also have contributed to the public benefit. But in a great many experiments (summarized in table 1) this is not the case.

As a benchmark, I will begin with a case of crowding in due to institutional complementarity. To explore the effects of explicit incentives in the laboratory, Gaechter, Kessler, and Konigstein (2004) implemented a "gift exchange game" (Fehr, Gachter, and Kirchsteiger (1997)) in which Swiss student subjects in the role of principals (employers) make a wage offer with a stipulated desired level of effort on the part of the agent (worker). The agent may then choose an effort level, with costs to the agent rising in effort. In the 'stranger treatment' the pairs were shuffled every period, so that each period was a one-shot interaction; the participants were certain they would not encounter any partner more than once. The best response for a self-regarding subjects in this treatment is for agents to provide minimal effort (one unit) irrespective of the wage, and for principals, inducting this, to offer the minimal wage. In the 'partner treatment' the two remained paired over ten periods, and this set of ten periods was itself repeated three times. Because the interaction was repeated with the same partner, subjects with self-regarding preferences in this treatment have reasons to provide higher wages and effort than the minimum, even if they believe their partner also to be self-regarding.

As in earlier experiments with this game, 'employers' made wage offers far more generous than the minimum required to elicit the minimum one unit of effort in the stranger treatment . The effort offered in return is much higher (four times higher) than would have been optimal for a self-regarding 'employee' so we can conclude that social preferences of

[3] Framing does not imply preference endogeneity. Framing makes behavior situationally dependent, but it is consistent with time-invariant situationally specific behavior: over time, one acts the same way in the same situation. By contrast, preferences are endogenous if one's experiences result in durable changes in behavior in given situations.

some sort were at work. Repeated interaction resulted in much higher levels of effort than the stranger treatment, and effort rose over the three sets of ten periods and also (except for a dramatic end of game drop off) within the sets of play. The fact that repetition contributed to cooperation, as well as the sharp effort reduction during the last two periods of play show that self interested incentives were effective. (The end game fall off is not likely due to learning as the decline is not monotonic within the sets and high (indeed higher) levels of effort are restored when the second and third sets are initiated). But the repeated interaction did more than to activate self-regarding motives: the reciprocal response to generous wages was 60 percent greater in the repeated treatment than in the one shot. The fact that end of set effort did not fall to the level of the stranger treatment also suggests that while repetition engaged the self-regarding motives it also tapped social preferences that the stranger treatment did not evoke.

Crowding in was also observed among experimental subjects from the University of Hokkaido who played a single-shot public goods game under three conditions: no punishment, punishment for low contributions of all group members, and a third condition designed to test crowding in by explicit incentives. In this condition one member of the group was liable to punishment for low contribution, but this was not known to other subjects, and the subject knew that no other subject would be punished and that the other subjects were not aware of the possibility of punishment. Yamagishi and Shinada (2006) found that mean contributions in the second condition (all subject to punishment) exceeded the no punishment mean contributions by 81 percent. In the third condition (one subject alone subject to punishment) contributions by the subject liable to punishment exceeded the no punishment level by 49 percent. The difference in the contribution levels in the second and third treatments is plausibly interpreted as the indirect effect of punishment in assuring the subject that others would contribute. The difference is entirely explained by differences in expectations (recorded in post experiment interviews) that the others would cooperate. Because contributing nothing is the dominant strategy for a subject with self-regarding preferences, the substantial indirect effect indicates that the knowledge that others would be punished for non contributions and the expectation that they would therefore contribute crowded in the subject's social preferences.

In this case the provision of incentives for self-regarding subjects improved performance and did not degrade (even enhanced) other-regarding preferences. But this is not generally the case. Fehr and Gaechter (2000) implemented the gift exchange game described above. In their "trust" treatment, the interaction ends when the agent chooses an effort level, as in the stranger treatment above. In the "incentive" treatment, following the agent's choice of an effort level, the employer may fine the worker, presumably using this option if the worker's effort level is thought to be inadequate. By contrast with the trust treatment, the incentive treatment links pay to performance and hence represents a more complete contract. In this experiment, the total surplus from the interaction is the principal's profits plus the

agent's wage minus the cost of effort (and the fine where applicable.)

As above, in the trust treatment, a self-regarding agent would choose the minimum feasible level of effort irrespective of the principal's wage offer, and, anticipating this, a self-regarding principal would offer the minimum wage. As in other experiments of this type, subjects did not conform to this expectation: Employers made generous offers and workers' effort levels were strongly conditioned on these offers, high wages being reciprocated by high levels of effort. The introduction of explicit incentives, however, had a negative effect: average effort levels by agents were substantially *lower*. The separability assumption failed in this case because under the incentive (fine) treatment, initially generous offers by employers were not reciprocated by higher employee effort, and once employers understood this, they made low offers. Thus the explicit incentive (the threat of the fine) appears to have reduced reciprocal motivations.

Inequality aversion among the agents may also have been involved. The experiment was constructed so that had subjects responded optimally on the basis of self-regarding preferences, the surplus would have been more than twice as great under the incentive treatment as under the trust treatment. But the total surplus was higher in the trust treatment, by 20 percent in those cases where the principal offered a contract such that the expected fine for shirking exceeded the cost of working (so that the no shirking condition was fulfilled), and by 53 percent where the principal's contract did not meet the no shirking condition.

An important result of this experiment emerges if we compare the distribution of the surplus under the trust treatment and the incentive treatment. In the incentive treatment (confining our attention to the cases in which the principal's contract fulfilled the no shirking condition) profits are more than double the profits in the trust treatment, while the net payoffs to the workers are less than half. The incentive treatment allowed employers to save enough in wage costs to offset the reductions in work effort. Summarizing this result, the authors write: "the incentive opportunities in the incentive treatment allow principals to increase their profits relative to the trust treatment, but ...this is associated with an efficiency loss."

Perverse incentive effects also occurred in a field experiment in Colombia (Cardenas, Stranlund, and Willis (2000)).The experiment captured the logic of a common pool resource extraction problem (over-exploitation of local forests) faced by the rural people who participated. In the absence of explicit incentives the subjects selected extraction levels not far above the social optimum and much less than what would have been the Nash equilibrium level assuming individual optimization with self-regarding preferences. But when monitoring of the subjects' extraction levels (by the experimenter) and the prospect of a fine for over-extraction were introduced, subjects extracted more rather than less. After a few rounds, their extraction levels approximated the new (self-regarding) Nash equilibrium level (taking account of the fine). The subjects apparently had switched from other-regarding to self-regarding

behavior as a result of the imposition of punishment. Like the fine imposed on the tardy Haifa parents, the effect of "improving" the incentive structure apparently was to diminish the salience of the other-regarding motives that had been in force in the absence of the incentives.

A related experiment may provide some insight into how and why the separability assumption fails (Frohlich and Oppenheimer (1995).) Subjects played 5-person public goods games under two conditions: one group played the standard contribution game and the other played a modified ('veil of ignorance') game in which a randomized assignment of payoffs made it optimal to contribute the maximal amount to the public good. Half of the subjects (in each treatment) were allowed to engage in discussion prior to each play (of course the discussion should have had no effect on the outcome of the standard game, as the dominant strategy is to contribute nothing). After 8 rounds of play, another 8 rounds were conducted, this time with the same groups but with all playing the standard game. Among those who had been permitted discussion, those who had experienced the incentive-compatible (veil of ignorance) game contributed significantly less in the final 8 rounds, and (in subsequent questionnaires) expressed less concern with questions of fairness.

The authors' explanation is that the incentive-compatible mechanism rewarded those contributing to the public good, thus making self-interest a good guide to action, while those experiencing the standard game gained high payoffs only to the extent that they evoked considerations of fairness as a distinct motive among their group-mates. They conclude

> The failure of the ... (incentive compatible) mechanism to confront subjects with an ethical dilemma appears to lead to little or no learning in ethical behavior in the subsequent period. ... It is an institution, like other incentive compatible devices, which can generate near optimal outcomes. ... However from an ethical point of view it is not only unsuccessful as pertains to subsequent behavior; it appears to be actually pernicious. It undermines ethical reasoning and ethically motivated behavior. (Frohlich and Oppenheimer (1995):44)

This interpretation is consistent with a large literature on the effects of performing various kinds of tasks on subsequent (sometimes seemingly unrelated) values (Breer and Locke (1965).) Other experiments have documented these dynamic crowding out effects (Irlenbusch and Sliwka (2004), Gaechter, Kessler, and Konigstein (2004)). In these two experiments, as in the case of the fines for tardiness at the Haifa day care centers, the negative effects of incentives persisted even after the incentives are no longer operative.

Fehr and List (2004) offered a different interpretation of counter- productive incentives found in their trust experiments with Costa Rican businessmen and students. The highest level of trustworthiness was elicited when the principal was *permitted* to fine the agent for

untrustworthy behavior, but had *pre-committed not to use it*, evidently a signal by the principal of trusting behavior that was then reciprocated by the agent. By contrast "explicit threats to penalize shirking, backfire by inducing less trustworthy behavior." They conclude: "the psychological message that is conveyed by incentives – whether they are perceived as kind or hostile – has important behavioral effect." Subjects in the identical experiments of Fehr and Rockenbach (2003) exhibited the same behavior. Trustees in the trust experiments by Falk and Kosfeld (2005) acted less trustworthy (they returned less of the Truster's transfer) when the Truster opted to impose a minimum return rate. In post-play interviews, most agreed with the statement that the imposition of the minimum was a signal of distrust; among the 57 per cent of Trustees who had reduced their transfer after the imposition of the minimum, this view was virtually unanimous (93 percent of them agreed with the statement).

Fines or other negative incentives may have strongly positive effects, however. We know from public goods with punishment experiments (Fehr and Gachter (2000), Ostrom, Walker, and Gardner (1992), Yamagishi (1988)) that defecting members of a group substantially increase their contributions if other members have paid to reduce the defector's payoffs. Carpenter, Bowles, and Gintis (2006) show that punishment is effective even when it is not sufficient to make positive contributions a best response (defined over the game payoffs) and Barr (2001) and Masclet, Noussair, Tucker, et al. (2003) find that the simple expression of disapproval by fellow members without any material punishment is effective. However, when (as occasionally occurs) high contributing members are punished by peers, they reduce their contributions in subsequent rounds(Carpenter, Bowles, and Gintis (2006)). These results are consistent with the view that negative incentives in the form of expressed disapproval (with or without payoff consequences) may evoke shame (if the subject feels guilty about his contribution) and other aspects of preferences not captured in the game payoffs. In his case negative incentives may 'crowd in' other-regarding motives. However when the target of punishment does not feel guilt, the result of punishment is spite rather than shame.

Experiments (mostly by psychologists) have identified conditions under which extrinsic rewards such as monetary payment for performance of a task diminish one's intrinsic motivation to do the task (Deci, Koestner, and Ryan (1999)). While these experiments continue to generate controversy (Cameron, Banko, and Pierce (2001), Eisenberger and Cameron (1996)), my reading of the evidence is that these crowding out effects occur when the relevant tasks are interesting rather than boring and when the reward is expected in advance and closely tied to the task performance. One may conclude that performance-based pay in work places may diminish employee's motivation to do tasks which they initially found intrinsically interesting or challenging. But the evidence is also consistent with an important role for explicit (extrinsic) incentives in motivating individuals to do tasks in which they have little intrinsic interest (that is to say, a great many jobs).

These extrinsic reward experiments differ from most economic experiments in two relevant ways. First, the public goods, gift exchange, and other games favored by economists are structured so that a purely self-regarding individual will contribute the minimal amount permitted; the finding of interest is that most experimental subjects do not behave this way. The intrinsic motivation experiments by psychologists consider activities that the subjects initially enjoy doing (painting pictures, for example) and show that pay for performance may degrade these initial positive motivations. Second, the incentives (extrinsic rewards) in the psychological experiments are typically implemented by the experimenter, not as in many of the economics experiments, by one or more of the strategically interacting subjects (as the gift exchange or other principal agent games). Thus the extrinsic incentives are not viewed as a signal of the type or intent by another subject.

Additional evidence of non-separability is found in other experiments, some of which are summarized in Table 1 (see also Frey and Jegen (2003)). I have not listed here the substantial literature on the 'crowding out of intrinsic by extrinsic motives' as that is adequately surveyed in the works cited above.

Drawing general conclusions from these experiments is difficult. For example, Fischbacher, Fong, and Fehr (2005) found that while low offers by proposers in a dyadic bargaining (Ultimatum) game are often rejected by respondents, this occurs much less frequently when there is competition among respondents. This result could be interpreted as showing that market competition crowds out fair-mindedness by depriving subjects of the expectation that their refusal of low offer will inflict a penalty on the unfair proposer. But the Henrich, Boyd, Bowles, et al. (2005) study of 15 small-scale societies could be interpreted as showing the opposite: in that study the likelihood that low ultimatum game offers are rejected was significantly greater in the more market-integrated societies. The experiments thus do not allow any simple interpretation. But a few lessons may be suggested.

4. *Optimal incentives in the absence of separability*

When separability does not hold, how are optimal incentives affected? Because crowding out reduces the effectiveness of explicit incentives, one might anticipate that their optimal level would be reduced, by comparison to the benchmark of separability (and conversely that crowding in would favor greater use of incentives). However if the incentive is less effective, the equilibrium allocational distortion sustained by a given tax rate will be larger, so greater use of the incentive might be warranted. Thus crowding out may raise both the marginal cost and the marginal benefit of the incentive. As we will see, incentives may be either overused or underused when crowding out is not taken into account; and the same conclusion holds for crowding in.

We model a two-stage optimization process in which a social planner selects a tax to

maximize the net benefits of a public good provision project, given the citizen's individual best responses to the incentives implemented by the tax (assumed known to the planner). Returning to the public goods problem in section 2, suppose that the "value utility" of contributing is given by $v = a(\underline{v} + \delta t)$ so $v_a = \underline{v} + \delta t$, where as before $\delta < 0$ represents crowding out, and conversely. Thus from the member's best response function (2) the effect of the tax on the individual's contribution is $a^*_t = 1 + \delta$ from which we see that if $\delta < -1$, raising taxes reduces contributions. This effect, consistent with the Haifa day care case and many other experiments, is termed *strong crowding out*.

We now take account of the costs of imposing the tax equal to $\tau t^2/2$ per member of the community. We assume that the planner's social welfare function does not take account of the value utility enjoyed by the members. Thus the benefits of a tax increase include the resulting increased provision of the public good net of the cost to individuals of providing it, but not the increased value satisfaction experienced by the citizens who responded positively to the tax. And (because it will not affect the results) we abstract from benefits associated with the tax revenues. Thus the planner varies t to maximize the net benefits of the project produced per member:

(3) $\omega(t) = - a^{*2}/2 + \phi n a^* - \tau t^2/2$

The optimal incentive thus implements

(4) $(n\phi - a^*) = \tau t^*/(1 + \delta)$

where the left hand expression is the marginal benefit of contributing net of the marginal cost of provision and the expression on the right the marginal effective tax cost (that is, the cost of the tax per unit of effect on contributions). Using (2) and rearranging (4) we have

(5) $n\phi - \{\phi + \underline{v} + (1 + \delta)t^*\} = \tau t^*/(1 + \delta).$

The fact that crowding affects the marginal net benefits of contributing and the marginal effective cost of the tax in the same direction explains the ambiguity of the effect of crowding on optimal incentives mentioned above.

The optimal incentive is given by

(6) $t^* = \{(\phi(n-1) - \underline{v}\}(1+\delta)/\{\tau + (1+\delta)^2\}$

for $\{\phi(n-1) - \underline{v}\} > 0$ and $(1+\delta) \geq 0$ which we assume throughout, that is, when ethical values alone are insufficient to internalize the external benefit of contributing when $t = 0$ and in the absence of strong crowding out (the latter assumption is sufficient for $\omega(t)$ to be concave

assuring that (4) is a maximum) . Otherwise $t^* = 0$. One may confirm that no optimal tax exists in the presence of strong crowding out ($\delta < -1$). We also have

(7) $\quad dt^*/d\delta = \{(\phi(n-1) - \underline{v}\}(\tau - (1+\delta)^2)/\{\tau + (1+\delta)^2\}^2$

so $\text{sgn}\{dt^*/d\delta\} = \text{sgn}((\tau - (1+\delta)^2)$

The optimal tax under separability is

$t^s = \{(n-1)\phi - \underline{v}\}/(\tau + 1)$.

This expression shows that under separability values are a substitute for incentives: larger $\underline{v}$ entails lower t^*. We say that incentives are overused if $t^s > t^*$ and conversely. The relationship between t^s and $t^*$is given by

$$t^s - t^* = \{(\phi(n-1) - \underline{v}\}[1/(\tau+1) - (1+\delta)/\{\tau + (1+\delta)^2\}]$$

so overuse occurs when

(8) $\quad t^s > t^* \Leftrightarrow 1/(\tau+1) > (1+\delta)/\{\tau + (1+\delta)^2\} \Leftrightarrow \delta\{\tau - (1+\delta)\} < 0$

If $\tau = (1+\delta)$ then $t^s = t^*$, while overuse of incentives occurs if

(Crowding in): $\delta > 0$ and $\tau < 1+ \delta$ and

(Crowding out): $\delta < 0$ and $\tau > 1 + \delta$

So crowding in and small marginal tax costs lead to the overuse of incentives, while the same is true of crowding out and substantial tax costs.

Figure 1 illustrates these relationships**.** Figure 2 illustrates the optimal tax and its dependence on δ and τ.

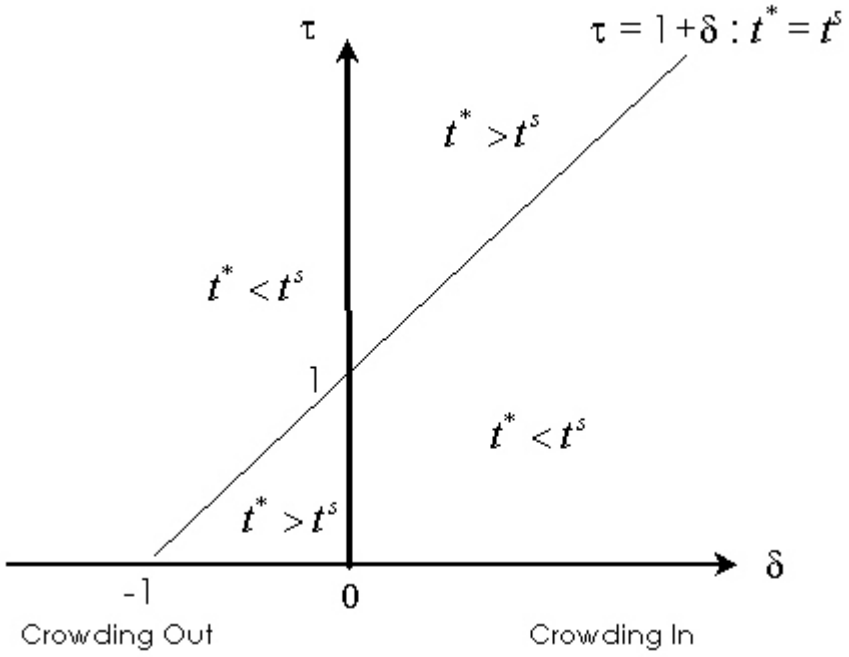

Figure 1. When are explicit incentives overused if non-separability is ignored?

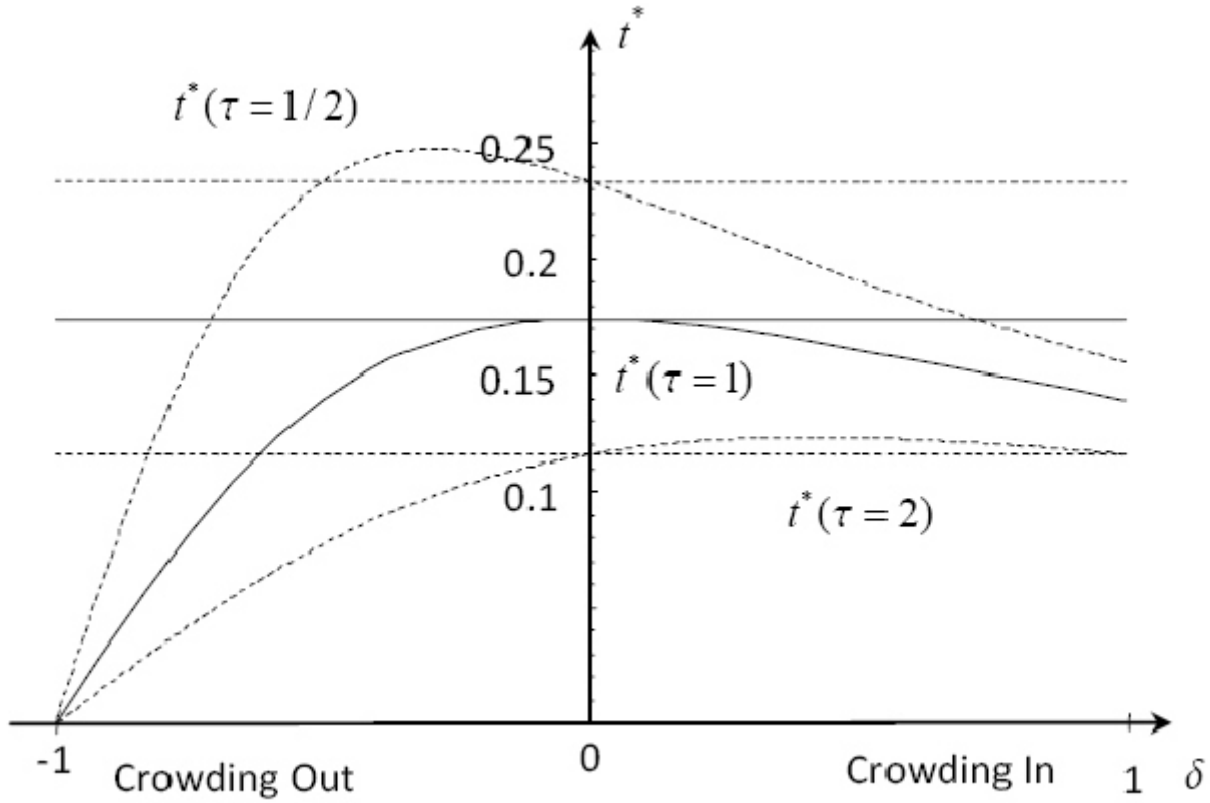

Figure 2. Optimal incentives under non-separability. Note: n = 10, ϕ = 0.15, $\underline{v}$ = 1

The vertical intercept in Figure 2 (that is for $\delta = 0$) gives t^s. When $\tau = 1$ and $\delta \neq 0$ incentives are overused, irrespective of whether crowding in or crowding out obtains: $t^s > t^*$ for all $\delta \neq 0$ because from (7) the maximum t^* occurs when $\tau = (1 + \delta)^2$. The economic intuition behind this curious result (evident from equation (6)) is that if $\tau = 1$ crowding out ($\delta < 0$) raises the effective marginal cost of the tax by $\tau t/(1+\delta)^2$, while it raises the net benefits of contribution (by reducing the marginal cost of contribution) by a smaller amount t, so lesser use of the incentive is optimal. By contrast, crowding in raises the marginal provision cost to the individual (and hence reduces marginal net benefits) more than it lowers the marginal effective cost of the tax. Thus in both cases the optimal tax is reduced.

5. Conclusion. Public economics in light of behavioral economics

Incentives work. This is particularly true of positive incentives and applies also to negative incentives that avoid conveying negative information about the type or intentions of those with whom the individual is interacting. In some experiments, the response to variations in a given incentive structure (variations in a piece rate or gain share, for example) closely approximates what one would expect based on self-regarding preferences (for example, Anderhub, Gaechter, and Konigstein (2000) Irlenbusch and Sliwka (2004)), consistent with the separability assumption. But the experimental evidence surveyed here also suggests that the effects of public spirited motives may be either enhanced or diminished by policy interventions designed to more closely align self-regarding incentives with social objectives.

But if incentives work, why should one be concerned about failures of the separability assumption? As long as there exists *some* combination of explicit incentives that will implement a social optimum without the assistance of other-regarding preferences, are these additional considerations not simply an unnecessary complication of normative public economics?

There are five reasons why the answer is "no". First in the presence of strong crowding out the premise of the question is false, as the incentives are literally counterproductive (their effect has the wrong sign) so the optimal explicit incentive will be zero. Second, the information necessary to implement an optimal fiat and contract allocation (supposing one to exist) would typically be unavailable to states and private principals, or extraordinarily expensive to acquire and use. Third, optimal incentives would be infeasible in many cases due to wealth constraints, the correction of which would pose insurmountable additional incentive problems. Fourth, those entrusted with designing and implementing optimal incentives would themselves need the proper incentives so that, as Bentham put it, their "interests" would coincide with their "duties." Fifth, one may value social preferences for reasons other than their contribution to allocational efficiency.

A more modest approach would be to recognize that explicit incentives do a tolerably good job in many situations and that in others their performance would be improved if their design took account of effects on social preferences. Social preferences are a fragile resource for the policy maker, one that may be either empowered by legislation and public policy, or diminished. This suggests an extension of Hume's maxim: Good policies and constitutions are those that support socially valued ends not only by harnessing selfish preferences, but also by evoking, cultivating and empowering public-spirited motives. This will be particularly important where contracts are incomplete; for it is in these cases that as Arrow (1971):22 put it: "norms of social behavior, including ethical and moral codes (may) ...compensate for market failures."

Where this is the case, as we have seenconventional incentive-based interventions may be worse than ineffective, motivating a norm-related analogue to the second best theorem due to Lipsey and Lancaster (1956-1957): where contracts are incomplete (and hence socially beneficial values may be important in attenuating market failures), public policies and legal practices that more closely approximate idealized complete contracting may exacerbate the underlying market failure (by undermining social values such as trust or reciprocity) and may result in a less efficient equilibrium allocation. A constitution for knaves, Bruno Frey (1997) observed, may produce knaves, just as Michael Taylor (1976) had earlier suggested that the Hobbesian state may produce Hobbesian man.

Table 1. Explicit Incentives and Social Preferences: Experiments

Citation	*Subject pool*	*Game*	*Result*	*Comment*
Gneezy and Rustichini (2000a)	Haifa daycare parents	Fine imposed for lateness	...increased lateness which persisted after fine was withdrawn	fine signaled 'how bad' lateness was, shifted 'from a communal to an exchange' relationship
Fehr and Gaechter (2002)	Swiss students	Gift Exchange (GE)	Explicit incentives reduce effort (especially if negative), redistribute surplus to principal.	Framing and inequality aversion Incentives eliminate the positive effects of generosity (31)
Upton (1974)	U.S blood donors	Paid donations or uncompensated	Highly motivated givers respond negatively to incentives	Substantiates Titmuss (1971) See: Bliss (1972), Arrow (1972)
Gneezy (2003)	U.S students	Proposer-Responder	W-curve: Non-monotonic effects of fines and rewards.	Discontinuity at zero reflects shift from moral to a strategic mode? See Gneezy and Rustichini (2000b)
Fehr and List (2004)	CEO's & students (Costa Rica)	Trust Game (TG) with optional punishment	Not using the punishment option when it is available results in high performance	Key: "the psychological message .. conveyed by incentives – whether ... kind or hostile..."

Bohnet, Frey, and Huck (2001)	U.S. students	Contract enforcement	Compliance is non-monotonic in degree of enforcement	"Monetary" crowd out "Honest" preferences where enforcement is moderately strong
Fehr and Rockenbach (2003)	German students	Trust Game with optional punishment	Not using the punishment option when it is available results in high performance	Forgoing the punishment option is a signal of good will and trust
Cardenas, Stranlund, and Willis (2000)	Colombian rural poor	Common pool resource with fines	Fines induce more self-interested behavior & pool over exploitation	Fine induced a shift from moral to self interested frame ?
Schotter, Weiss, and Zapater (1996)	U.S. students	Ultimatum and Dictator Games	competitive threats to survival induced lower offers	"..[market] offers justifications for actions that in isolation would be unjustifiable" p.38
Fehr, Gachter, and Kirchsteiger (1997)	Swiss students	Gift Exchange (effort non-contractible)	Monitoring and fines reduced effort	
Frohlich and Oppenheimer (1995)	Canadian students	Prisoners' Dilemma (PD)	Incentive compatible option reduced performance in subsequent play	IC option 'undermines ethical reasoning and ethically motivated behavior.' p.44
Gneezy and Rustichini (2000b)	Israeli students	Payment for soliciting contributions to social causes	Payment may reduce the performance of the solicitors	

Falk and Kosfeld (2005)	Swiss students	Trust Game	Ps who impose a minimum return rate on trustees receive less than trusting Ps	imposed minimum understood by S's as a sign of distrust by Ps
Hauser, Xiao, McCabe, et al. (2004)	U.S. students	Trust Game	Weak sanctions by Truster or by Nature induce less 'trustworthiness' .	"Extrinsic incentives ...can ...change subjects' frame from ethical to income-maximizing."
Fischbacher, Fong, and Fehr (2005)	Swiss students	"Bargaining" vs "Market" Ultimatum Game	Competition among respondents reduced rejections	Competition made punishment of 'unfair' offers less certain
Bohnet and Baytelman (2005)	senior executives in U.S.	TG: one shot, repeated, w/o & w punishment, communication ("institutions")	institutionals increase amount sent and (conditional on that) returned; option of punishment reduces offers of other-regarding trustees	"punishment [option] destroys intrinsic trust and...controlling for expectations of trust, lowers..willingness to reward trust"
Henrich, Boyd, Bowles, et al. (2005)	15 small scale societies	Ultimatum Game	Offers and rejection of low offers were greater in more market-integrated societies	"*doux commerce*"? Hirschman (1977)
Fehr, Klein, and Schmidt (2001)	German students	Gift exchange with piece rate and incomplete contracts	Incomplete (bonus) contracts yield higher returns to both P and A and are more common.	'existence of fairminded As may [explain] why many contracts are ...left incomplete'

Galbiati and Vertova (2005)	Italian students	Public goods game with rewards and penalties	stated contribution norm raises contributions independently of self-regarding incentives.	Contributions respond to socially determined 'obligations' (crowding in)
Tyran and Feld (2004)	Swiss students	Public goods with mild and strong sanctions	'compliance is much improved if mild law is endogenously chosen i.e. self imposed'	self imposed punishment does not indicate hostile intent
Gaechter, Kessler, and Konigstein (2004)	Swiss students	Gift exchange with fine, bonus, and trust	Cooperation is reduced in rounds subsequent to an incentive treatment; larger effect for fine than bonus	"Irreversibility: .. Incentives have a lasting negative effect on voluntary cooperation"
Hoffman, McCabe, Shachat, et al. (1994)	U.S. students	Ultimatum game	Market 'labels' (Exchange game) reduced offers and raised acceptance levels	Market framing induces self-regarding preferences
Irlenbusch and Sliwka (2004)	German students (Erfurt)	Gift exchange (wage-effort) with piece rate option	Piece rates lower effort when they are in force, and after they are abandoned.	"..incentive [suggests] an individual maximization frame rather than a cooperative frame"
Rustrom (2002)	U.S. students	Creative task ('tower of Hanoi') with large, small and no penalties and rewards	Penalties degraded performance; large rewards induced better performance than small (but no better than the no-incentive treatment)	Penalties 'distracted' S's

Tenbrunsel and Messick (1999)	U.S. students	social dilemma with weak, strong and no sanctions	Ss evaluated sanction treatment as 'business' rather than 'ethical' Weak sanctions decreased expectations others would cooperate.	Weak (strong) sanctions reduce (increase) cooperation; no effect of sanctions for those adopting an ethical frame
Gaechter and Falk (2002)	Austrian students	One shot and repeated gift exchange game	Reciprocity stronger in repeated game; repetition induces selfish agents to imitate reciprocators	Repetition does not reduce reciprocal motives and "crowds in" 'imitated' reciprocity
Carpenter, Bowles, and Gintis (2006)	U.S. students	Public goods with punishment	Peer punishment induced defectors to contribute more, even when defection remained a best response	Punishment activated guilt, crowding in shame induced cooperation.
Falk, Fehr, and Zehnder (2006)	Swiss Students	Labor market game with minimum wages	Minimum wages permanently raised reservation wages (even after the min wage ended)	"Min wages affect [subjects'] fairness perceptions" creating moral "entitlements"

Note: P is principal, S is subject.

Works cited

Anderhub, Vital, Simon Gaechter, and Manfred Konigstein. 2000. "Efficient Contracting and Fair lay in a Simple Principal Agent Experiment." *Institute for Empirical Research in Economics*.

Andreoni, James, Brian Erand, and Jonathan Feinstein. 1998. "Tax Compliance." *Journal of Economic Literature*, 36:2, pp. 818-60.

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

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

Barr, Abbigail. 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.

Benabou, Roland and Jean Tirole. 2005. "Incentives and prosocial behavior."

Ben-Porath, Yoram. 1980. "The F-Connection: Families, Friends, and Firms and the Organization of Exchange." *Population and Development Review*, 6:1, pp. 1-30.

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

Bohnet, Iris and Yael Baytelman. 2005. "Institutions and Trust: Implications for Preferences, Beliefs and Behavior."

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

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

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

111.

Breer, Paul E. and Edwin A. Locke. 1965. *Task experience as a source of attitudes*. Homewood, Ill.,: Dorsey Press.

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

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

Carpenter, Jeffrey, Samuel Bowles, and Herbert Gintis. 2006. "Mutual Monitoring in Teams: The Importance of Shame and Punishment." *University of Massachusetts*.

Cooter, Robert. 1998. "Expressive Law and Economics." *Journal of Legal Studies*, 27, pp. 585-608.

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

Deci, Edward L., Richard Koestner, and Richard 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.

Eisenberger, R and J Cameron. 1996. "Detrimental effects of reward: reality or myth." *American Psychologist*, 51, pp. 1153-66.

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

Falk, Armin and Michael Kosfeld. 2005. "Distrust: the hidden cost of incentives." *University of Bonn*.

Farooq, Omer. 2005. "Drumming tax sense into evaders." BBC News.

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

Fehr, Ernst, Simon Gachter, and Georg Kirchsteiger. 1997. "Reciprocity as a Contract

Enforcement Device: Experimental Evidence." *Econometrica*, 65:4, pp. 833-60.

Fehr, Ernst and Simon Gaechter. 2000. "Fairness and Retaliation: The Economics of Reciprocity." *Journal of Economic Perspectives*, 14:3, pp. 159-81.

Fehr, Ernst and Simon Gaechter. 2002. "Do Incentive Contracts Crowd Out Voluntary Cooperation?" University of Zurich.

Fehr, Ernst, Alexander Klein, and Klaus M. Schmidt. 2001. "Fairness, Incentives and Contractual Incompleteness." *CESifo and CEPR*.

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

Fischbacher, Uris, Christina Fong, and Ernst Fehr. 2005. "Fairness, errors, and the power of competition."

Fong, Christina, Samuel Bowles, and Herbert Gintis. 2005. "Strong reciprocity and the welfare state," in *Handbook of Giving, Reciprocity, and Altrusim*. Sege-Christophe Kolm and Jean Mercier Ythier eds. Amsterdam: Elsevier.

Frey, B. and R Jegen. 2003. "Motivation Crowding Theory: A Survey of Empriical Evidence." *Journal of Economic Surveys*, 15:5, pp. 589 - 611.

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

Frohlich, Norman and Joe A. Oppenheimer. 1995. "The Incompatibility of Incentive Compatible Devices and Ethical Behavior: Some Experimental Results and Insights." *Public Choice Studies*, 25, pp. 24-51.

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

Gaechter, Simon, Esther Kessler, and Manfred Konigstein. 2004. "Performance Incentives and the Dynamics of Voluntary Cooperation."

Galbiati, Roberto and Pietro Vertova. 2005. "Law and Behavior in Social Dilemmas." *University of Siena*.

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

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

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

Hauser, Daniel, Erte Xiao, Kevin McCabe, and Vernon Smith. 2004. "When punishment fails: Research on sanctions, intentions, and non cooperation." *George Mason University*.

Henrich, Joe, Robert Boyd, Samuel Bowles, Colin Camerer, Ernst Fehr, Herbert Gintis, Richard McElreath, Michael Alvard, Abigail Barr, Jean Ensminger, Natalie Smith Henrich, Kim Hill, Francisco Gil-White, Michael Gurven, Frank Marlowe, John Patton, and David Tracer. 2005. "'Economic Man' in Cross-Cultural Perspective: Behavioral experiments in 15 small-scale societies." *Behavioral and Brain Sciences*, 28.

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

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

Hoffman, Elizabeth, Kevin McCabe, Keith Shachat, and Vernon L. Smith. 1994. "Preferences, Property Rights, and Anonymity in Bargaining Games." *Games and Economic Behavior*, 7:3, pp. 346-80.

Hume, David. 1964. *David Hume, The Philosophical Works*. Darmstadt: Scientia Verlag Aalen.

Irlenbusch, Bernd and Dirk Sliwka. 2004. "Incentives, decision frames, and motivation crowding out: an experimental investigation." *London School of Economics*.

Kahan, Dan M. 1997. "Social Influence, Social Meaning, and Deterrence." *Virginia Law Review* (Virginia Law Review), 83:2, pp. 349- 95.

Kahneman, Daniel and Amos Tversky. 1986. "Rational Choice and the Framing of Decisions." *Journal of Business*, 59:4, pp. S251-78.

Kreps, David M. 1997. "Intrinsic motivation and extrinsic incentives." *American Economic Review*, 87, pp. 359-64.

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

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

Lipsey, R. and K. Lancaster. 1956-1957. "The General Theory of the Second Best." *Review of Economic Studies*, 24:1, pp. 11-32.

Masclet, David, Charles Noussair, Steven Tucker, and Marie-Claire Villeval. 2003. "Monetary and Non-monetary Punishment in the Voluntary Contributions Mechanism." *American Economic Review*, 93:1, pp. 366-80.

Ostrom, Elinor. 2000. "Crowding out Citizenship." *Scandinavian Political Studies*, 23:1, pp. 3-16.

Ostrom, Elinor, James Walker, and Roy Gardner. 1992. "Covenants with and without a Sword: Self-Governance Is Possible." *American Political Science Review*, 86:2, pp. 404-17.

Pommerehne, W.W. and Hannelore Weck-Hannermann. 1996. "Tax rates, tax administration and income tax evasion in Switzerland." *Public Choice*, 88:1-2, pp. 161-70.

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

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

Seabright, Paul. 2004. "Continuous Preferences Can Cause Discontinuous Choices: An application to the impact of incentives on altruism."

Sliwka, Dirk. 2007. "Trust as a Signal of a Social Norm and the Hidden Costs of Incentive Schemes." *American Economic Review*, in press.

Taylor, Michael. 1976. *Anarchy and Cooperation*. London: John Wiley and Sons.

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

Tilly, Charles. 1981. "Charivaris, Repertoires and Urban Politics," in *French Cities in the Nineteenth Century*. John M. Merriman ed. New York: Holmes and Meier, pp. 73-91.

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

Tyran, Jean-Robert and Lars Feld. 2004. "Achieving Compliance when Legal Sanctions are Non-deterrent."

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

Yamagishi, Toshio. 1988. "The Provision of a Sanctioning System in the United States and Japan." *Social Psychology Quarterly* (Social Psychology Quarterly), 51:3, pp. 265-71.

Yamagishi, Toshio and Mizuhu Shinada. 2006. "Punishing Free-riders: Direct and indirect promotion of cooperation." *Hokkaido University*.