Social Preferences and Public Economics: Mechanism design when social preferences depend on incentives

Samuel Bowles and Sung-Ha Hwang[1]

15 November, 2007

Abstract

Social preferences such as altruism, reciprocity, intrinsic motivation and a desire to uphold ethical norms are essential to good government, often facilitating socially desirable allocations that would be unattainable by incentives that appeal solely to self-interest. But experimental and other evidence indicates that the effect of conventional economic incentives and social preferences may be either complements or substitutes, explicit incentives crowding in or crowding out social preferences. We investigate the design of optimal incentives to contribute to a public good under these conditions. We identify cases in which a naive social planner will over-use or under-use explicit incentives by comparison to those that would be adopted by a sophisticated planner cognizant of these non-additive effects.

JEL: D52 (incomplete markets), D64 (altruism), H21 (efficiency, optimal taxation) H41, (public goods)

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

[1] Thanks to the Behavioral Sciences Program of the Santa Fe Institute, the National Science Foundation, and the University of Siena for financial support of this project, and the Certosa di Pontignano (Siena) for providing an ideal research environment. We would also like to thank Margaret Alexander, Ernst Fehr, Duncan Foley, John Geanakoplos, Suresh Naidu, Seung-Yun Oh, John Roemer, Bob Rowthorn, Paul Seabright, Rajiv Sethi, Joaquim Silvestre, Peter Skott, E. Somanathan, Tim Taylor, Elisabeth Wood, and Giulio Zanella for their contributions to this research. Affiliations: Bowles (corresponding author), Santa Fe Institute and University of Siena; Hwang, University of Massachusetts at Amherst.

1. Introduction

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)).

But Hume only "supposed" citizens to be knaves. In recent years experimental evidence has endorsed Hume's caveat (immediately following the above passage) that the maxim is "false in fact": preferences such as altruism, reciprocity, spite and intrinsic motivation are powerful and common motivations (Camerer (2003) Fehr, Klein, and Schmidt (2007), Gintis, et al. (2005)). The empirical importance of other-regarding motives for public economics has also long been recognized and has recently been affirmed in studies of tax compliance (Andreoni, Erand, and Feinstein (1998) Pommerehne and Weck-Hannemann (1996)), political opinion and voting concerning income security and redistribution measures (Fong, Bowles, and Gintis (2005)), and generalized obedience to law (Kahan (1997)).

Hume, Jeremy Bentham and the other classicals advocating self-interest as a basis of public policy design did not ignore the social preferences that underlie moral behavior. Instead they assumed that ethical motivations would be unaffected by incentive-based policies designed to harness self-interest. Along with civic virtue, thus explicit incentives and constraints and civic virtue could contribute additively to good government. According to this view, taxes or subsidies affect individual utility and hence behavior only by altering the economic costs and benefits of the targeted activities. These and other explicit incentives thus do not appear directly in the citizen's utility function. A consequence of the classicals' implicit 'separability assumption' is that 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 compromise or enhance civic virtue. Modern public economics, implementation theory, mechanism design and related fields continue this practice.

If only self-regarding motives were at work, the separability assumption could not fail. The reason is that the policy maker would then be working with a *tabla rasa*: the mobilization of private self-regarding motives to serve common public objectives could not extinguish other motives that might also have contributed to the public benefit. However a great many experiments and observations in natural settings suggest that other-regarding preferences are often important influences on behavior, and that the salience of these preferences varies with the kinds of explicit incentives that are implemented. Some of the experimental evidence is summarized in Table 1.

The underlying social and psychological mechanisms accounting for non-separability include the following. Explicit incentives may frame a decision setting as one in which self-interested optimization rather than ethical behavior is appropriate (Hoffman, McCabe, Shachat, et al. (1994), Irlenbusch and Sliwka (2005), Cardenas, Stranlund and Willis (2000), Gneezy and Rustichini (2000a)). Alternatively, the incentives adopted by a principal unavoidably provide information about the principal's preferences as well as his beliefs about the trustworthiness of the agent or other aspects of the agent's likely behavior The use of explicit incentives thus may convey distrust or other negative beliefs or attitudes by the principal towards the agent or may reveal that the principal would like to profit unfairly at the expense of the agent, thereby compromising the agent's preexisting predispositions of reciprocity or obligation toward the principal (Falk and Kosfeld (2006), Fehr and List (2004), Fehr and Rockenbach (2003)).

Further, rewards closely linked to performance may result in what psychologists term 'over-justification' which by compromising the individual's sense of self-determination may degrade intrinsic motives to perform well (Upton (1974), Deci, Koestner and Ryan (1999), Cameron and Pierce (2001), Kreps (1997), Frey (1994)). Moreover, the incentives adopted by a principal influence the process by which agents update their preference and may bias the endogenous formation of preferences in a self-interested direction (Bohnet, Frey and Huck (2001), Bowles (1998), Gaechter, Kessler, and Konigstein (2004), Frohlich and Oppenheimer (1995), Bar-Gill and Fershtman (2005)). Finally, explicit incentives may also crowd in ethical and other social preferences, as for example when members of a community prefer to contribute to a public good conditional on others contributing, and the presence of explicit incentives to contribute affects their beliefs about the actions likely to be taken by other members (Shinada and Yamagishi (2007), Sobel (2007), Fischbacher, Fong, and Fehr (2005), Fehr and Gaecher (2000), Rodriguez-Sickert, Guzman, and Cardenas (2007)).

If as these experiments suggest, the separability assumption is false, policies designed on its basis will generally be non-non-optimal, with explicit incentives being over-used or under-used. Over-use of explicit incentives when crowding out obtains was the central theme of the study of blood donations by Richard Titmuss (1971). In similar vein Albert Hirschman (1985):10 castigated economists who propose "to deal with unethical or anti-social behavior [solely] by raising the cost of that behavior…[because they] think of citizens as consumers with unchanging or arbitrarily changing tastes" adding that "A principal purpose of publicly proclaimed laws and regulations is to stigmatize antisocial behavior and thereby to influence citizens' values and behavioral codes." The implications for constitutional design of cases in which "institutions themselves affect preferences" were first developed by Michael Taylor (1987):177 and subsequently expanded by Bowles (1989), Frohlich and Oppenheimer (1995), Kreps (1997), Frey (1997), Cooter (1998), and Ostrom (2000), Bar-Gill and Fershtman (2005).

The economic intuition underlying these works is that because crowding out reduces their effectiveness, explicit incentive should be used less, and will be overused by a naive social planner who assumes that economic and moral motives are separable. If crowding out is so strong that the incentive is literally counter productive, having an effect the opposite of its intent, this is of course the case. But in general the effect of crowding out on the optimal use of incentives is far from obvious. The reduced effectiveness of the incentive associated with crowding out would entail a larger incentive for a planner designing a subsidy to ensure compliance with a quantitative target, a given fraction of the population receiving anti-flu injections for example. We will show that these seemingly conflicting intuitions are both correct by developing a model of optimal explicit incentives in the presence of both crowding in and crowding out, and using the model to identify cases in which crowding out entails greater or lesser use of incentives and conversely.

To analyze these cases we will ask what incentives would be adopted by a social planner who wishes to maximize the aggregate utility of a group of citizens. We will say that incentives are over-used if the sophisticated planner who takes account on non-separability would adopt a lesser level of incentive than would the naive planner, and conversely.

What does it mean to maximize total utility when ethical or other preferences may induce citizens to sacrifice personal pleasures in order to uphold moral norms? Bentham distinguished between the citizens' "interests" and their "duties;" and while advocating public policies to more closely align the two, he did not equate them from a normative standpoint. Ought the planner to reject Bentham's distinction and consider the satisfaction of the citizen's ethical values hedonistically, treating their contributions to publically valuable projects as not different in kind from their consumption of goods and services? In this case ethical values are just another form of 'tastes,' the satisfaction of which is pleasurable to the citizen. Or, recognizing that ethical values may require citizens to act in ways that entail personal sacrifices, should the planner treat values as a component of individual motivation but not part of the social welfare function?

The problem is not specific to the case of ethical and other social preferences. It arises because individual utility functions play both a positive and a normative role in public economics – explaining individual behaviors and how they are effected by alternative public policies, and evaluating the consequences of the policies under study. The former may require taking account of addictions, hyperbolic discounting, weakness of will and other empirically observed aspects of motivation. In cases where these motives lead to destructive or irrational behavior, the case for not treating their satisfaction as part of the normative standard for policies is clear enough. By contrast, acting in conformity with one's ethical values may be a source of either enduring satisfaction or pain, so the appropriate treatment is ambiguous. Thus we explore optimal incentives for two classes of sophisticated planner. For the first, preferences are revealed only by behaviors, so we have no basis for distinguishing between values and other 'tastes.' The second restricts the effect of a project on public welfare to its conventionally defined benefits and costs. We call the first the revealed preference planner and the second the conventional planner.

In the next section we introduce a model of public incentives when individuals with social preferences may contribute to a public good and we use this model to clarify the separability assumption and how it may be violated. We use the model to show that the sophisticated social planner seeking to ensure a target compliance level of contributions by citizens will implement a higher level of incentives (or none at all) if crowding out obtains. We then address two additional cases. In section 3 we study optimal incentives for the revealed preference planner, finding that in contrast to the compliance case, incentives will be overused when crowding out obtains and conversely. In section 4 we study the conventional planner's optimal incentives, finding that under use of incentives by the naive planner may obtain under crowding in or (counter-intuitively) crowding out. In section 5 we consider some of the implications of non-separability for public economics.

2. Non-Separability and compliance with social targets

We abstract from the diverse reasons why separability may fail and simply attribute to citizens a set of 'values' that may motivate pro-social behaviors and let these values be influenced (positively or negatively) by the use of explicit incentives. Consider a community of identical individuals indexed by $i = 1,...n$ who may contribute to a public project by taking an action ($a_i \in [0,1]$) at a cost $g(a_i)$ which is increasing and convex in its argument. The output of the project depends on each member's contributions, $f(a_1, a_2, ..., a_n)$ and explicit incentives take the form of a subsidy $s > 0$ proportional to the amount contributed. Implementing the subsidy entails monitoring and other costs $c(s)$ that are increasing in the level of the subsidy because higher values of s increase the citizens' incentive to misrepresent the level of their contribution. We suppose that raising the revenue supporting the subsidy has no effect on the problem and can be ignored.

We refer to ethical, other-regarding and other social preference influences on behavior as 'values' and represent them by $v(a_i, s)$. For clarity we refer to the benefits and costs other than values (the cost of contributing and receiving and administering subsidies as well as the benefits of the project) as "material". To isolate the problem of non-separability we abstract from individual differences in the effects of incentives on values and give the values function an explicit form

$$v = a_i(\underline{v} + \delta s) \qquad (1)$$

so the marginal effect of contributing on values is $v_{a_i} = \underline{v} + \delta s$. Not all of the psychological mechanisms accounting for non-separability are captured by this simple formulation; for example plausible cases in which the presence of the incentive has a substantial effect on values even if the incentive is arbitrarily small or where the effect of incentives on values depends on the actions or values of others are precluded. However it illustrates the fundamental problem of values and incentives being either complements or substitutes. The classical separability assumption maintains that the level of explicit material incentives does not influence the marginal value utility of contributing: that is $\delta = 0$. Then individual i 's utility is

(2) $$u_i = f\left(a_1, a_2, .., a_n\right) + sa_i - g(a_i) + v(a_i, s)$$

Because we wish to model the under-provision of a public good under private incentives and the possible implementation of a superior outcome through a publically imposed incentive, we make the following assumptions

1. In the absence of subsidy the marginal benefits that one's contributions confer on the community exceed one's private marginal costs, which exceed private marginal benefits (both material and value); as a result, without incentives, the public good will be under-provided; for all i, for $a_i \in [0,1]$, $nf_{a_i} > g' > f_{a_i} + \underline{v}$.

2. The individual cannot experience a negative valuation of contributing unless strong crowding out obtains; i.e. $\underline{v} \geq s$, which insures that $v(a,s) \geq 0$ for all $\delta > -1$.

Using (1) and (2) the individual's best response a_i is given by

(3) $$g'(a_i) = f_{a_i} + s + \underline{v} + \delta s$$

where the left hand side is the private marginal material cost of contributing and the remaining (right hand side) terms are private marginal material benefits arising from the project and from subsidies and the marginal value benefits associated with the individual's values. The effect of the subsidy on the individual's contribution (given the contributions of others) is then

(4) $$\frac{\partial a_i}{\partial s} = \frac{1+\delta}{g'' - f_{a_i a_i}}$$

where the denominator is positive by the second order condition of the individual's optimization problem (the marginal costs of contributing must be rising faster than the marginal private benefits).

Where the separability condition does not hold, we have either crowding in ($\delta > 0$) or crowding out ($\delta < 0$). Under crowding in, values and incentives are complements, as increased use of the incentive enhances the marginal effect of contributing on one's values and by (4) increases the effect of the subsidy on the citizen's action. Crowding out makes incentives and values substitutes, reducing the effect of incentives on the citizens' behavior. If $\delta < -1$, which we term strong crowding out, the incentive reduces contributions.

To explore the effects of non-separability we first study a problem of securing compliance with a target level of citizen contributions. Suppose a social planner seeks to ensure at least cost that at least p percent of the population contribute some minimum, $\bar{a}$. For concreteness suppose the action is training in first aid, measured in hours, and a social planner knows that in the absence of a subsidy this will not occur. He is constrained not to discriminate among the citizens and so considers a subsidy s applied to each hour of training received by the citizens where $c(s)$ is the cost of determining the number of hours contributed by each. We suppose that the benefit function takes the following form.

(5) $$f(a_1, a_2, ..., a_n) = \sum_i \phi_i a_i$$

where ϕ_i is constant for each i as the general benefits of an individual having first aid knowledge differ across individuals. We reorder the index such that $\phi_i \leq \phi_j$ for $i < j$.

Then an individual's utility is

(6) $$u_i = \sum_i \phi_i a_i + s a_i - g(a_i) + a_i \underline{v} + a_i \delta s$$

Therefore the individual's best response is given by

(7) $$g' = \phi_i + s + \underline{v} + \delta s$$

We identify the marginal individual who must contribute $\bar{a}$ in order to secure the compliance target of the planner as $\bar{i}$ so $\bar{i}$ is the smallest number, i, satisfying $i > n(1-p)$.

Then the social planner will choose $s^*(\delta)=0$ if $\delta \le -1$ abandoning the target as unattainable by use of the subsidy, and otherwise select the subsidy satisfying

(8) $$g'(\overline{a}) \le \phi_{\bar{i}} + s^*(\delta) + \underline{v} + \delta s^*(\delta)$$

Since providing the subsidy is costly, the social planner will choose the minimum $s^*(\delta)$ satisfying (8) for $\delta > -1$.

(9) $$s^*(\delta) = \frac{g'(\overline{a}) - (\phi_{\bar{i}} + \underline{v})}{1+\delta}$$

Optimal incentives under the separability assumption are denoted, $s^s = s^*(0)$. The planner is naive if he falsely believes that, $\delta = 0$ and as a result adopts $s^s = g'(\overline{a}) - (\phi_{\bar{i}} + \underline{v})$ as his preferred subsidy. Then we say that incentives are under-used if $s^* > s^s$ and conversely. Since compliance with the target would not be secured without the subsidy ($g'(\overline{a}) > \phi_{\bar{i}} + \underline{v}$ assumption 1), (9) gives the following.

Proposition 1. To secure compliance to a given target, crowding out requires a larger incentive and crowding in a smaller one.

3. Optimal incentives for the revealed preference planner

We turn now to the problem in which the planner seeks to maximize the sum of citizens' utilities by adopting an optimal incentive in the presence of a public goods problem, in which the levels of contribution of each citizen affect the marginal benefits of other citizens' contributions. The output of the project varies with the sum of the contributions of the members and each member receives an amount:

(10) $$f(a_1, a_2, ..., a_n) = \phi\left(\sum_k a_k\right)$$

where ϕ is increasing in its argument.

We model a two-stage optimization process in which the planner selects a subsidy level to maximize citizens' utility, taking account of the effect of the subsidy on the citizens' Nash equilibrium contribution levels (assumed known to the planner.) We solve (3) for all $i = 1, \ldots n$ to find a Nash equilibrium given a subsidy s. Because citizens are identical and experience a rising marginal cost of contribution, the planner will implement a symmetric equilibrium. Thus we denote each individual's Nash equilibrium contribution as a^*, suppressing the individual subscript, which satisfies the following condition:

(11) $$g'(a^*) = \phi'\left(na^*\right) + s + \underline{v} + \delta s$$

Using (11) we can find the effect of the incentive on citizens' Nash equilibrium contributions.

(12) $$\frac{da^*}{ds} = \frac{1+\delta}{g'' - n\phi''}$$

where the asymptotic stability of the Nash equilibrium requires the denominator to be positive. Equation (12) differs from the individual best response function (4) because it takes account of the reciprocal influence of the actions of all other citizens on one's own incentives to contribute thereby capturing the full effect of the incentive in displacing the Nash equilibrium level of contributions. The effect of the subsidies is diminished if the benefit function is concave and multiplied if it is convex, as expected. Like equation (4) it shows that strong crowding out precludes the use of the incentive as the planner will adopt the incentive only if it affects citizen behavior in the intended direction.

We model the planner's problem with respect to a single citizen using the R superscript to designate the revealed preference planner, who thus varies s to maximize

(13) $$\omega^{\mathrm{R}}(s) = \phi(na^*(s)) - g(a^*(s)) + v(a^*(s), s) - c(s)$$

The optimal incentive satisfies

(14) $$s^*(\delta) = \arg\max_{s} \omega^{\mathrm{U}}(s)$$

The optimal incentive satisfies the following first order condition:

(15) $$(n\phi' - g'(a^*(s)) + \underline{v} + \delta s)\frac{da^*}{ds} + a^*\delta - c'(s) = 0$$

The first term in the left hand expression captures the net indirect effect of the change in contributions induced by variation in the subsidy, showing that the revealed preference planner takes account of the fact that the value benefits partially offset the material costs of contributing for the individual. The second term is the direct positive or negative effect of the incentive on values. The final term is the marginal administrative cost.

Using (11), the citizen's best response, we find that the marginal cost of contributing for the individual net of the marginal value benefits, namely, $g' - \underline{v} - \delta s$ is just $\phi' + s$. So using (12) and rearranging (15) we see that the optimal subsidy equates marginal benefits of the subsidy to its marginal costs:

(16) $$\underbrace{((n-1)\phi' - s)}_{\partial\omega^{\mathrm{R}}/\partial a^*}\ \frac{(1+\delta)}{\underbrace{g'' - n\phi''}_{da^*/ds}} = \underbrace{c'(s) - a^*(s)\delta}_{-\partial\omega^{\mathrm{R}}/\partial s}$$

In equation (16) we know that $\partial\omega^{\mathrm{R}}/\partial a^*$ is positive because $(n-1)\phi' > \underline{v} \geq s$ (assumptions 1 and 2) and as a result for lower values of δ (crowding out), the indirect effect on marginal net social benefits (the left hand side) is smaller. The marginal cost net of the subsidy effect on values (the right hand side) is greater. Thus because crowding out lowers marginal benefits and raises marginal costs, one would expect the optimal level of the subsidy to be less than s^s when δ is negative, and conversely. The following proposition confirms this intuition.

Proposition 2. Over-use of incentives by the naive revealed preference planner under crowding out and conversely. *We have the following cases:*

a) $s^*(\delta) = 0$ *for* $\delta \leq -1$*;*

b) $s^s > s^*(\delta)$ *for* $-1 < \delta < 0$ *; and*

c) $s^s < s^*(\delta)$ *for* $0 < \delta < 1$

Proof. Part a) follows directly from (12). b) and c) are implied by assumptions 1 and 2, see Appendix.

4. *Optimal incentives for the conventional planner*

The reason why crowding out entails lesser use of incentives for the revealed preference planner is that the subsidy suffers two liabilities: it is less effective (12) and its use reduces the utility of any citizen who is contributing a positive amount. This double jeopardy problem does not occur for the conventional planner. The conventional planner is aware of the citizens' ethical values and takes account of their effects on behavior, but his objectives are entirely conventional, varying s to maximize the following objective function.

(17) $$\omega^{\mathrm{C}}(s) = \phi(na^*(s)) - g(a^*(s)) - c(s)$$

The conventional planner's optimal subsidy is given by the first order condition.

(18) $$(n\phi' - g')\frac{1+\delta}{g'' - n\phi''} = c'(s),$$

As before using (11) to eliminate g', we have

(19) $$((n-1)\phi' - \underline{v} - (1+\delta)s)\frac{1+\delta}{g'' - n\phi''} = c'(s)$$

Comparing this to equation (16) we see that there are two differences. First, for the conventional planner the marginal cost of the subsidy is just $c'(s)$ rather than $c'(s) - a^*\delta$: the conventional planner does not take account of the direct (positive or negative) incentive effects on values. Second, the conventional planner considers just the marginal material costs g' (rather than

$g' - v_{a_i}$) as the marginal private cost to the individual of contributing. Because v_{a_i} is non-negative (by assumption 2) the marginal costs to the citizens of contributing are greater in the eyes of the conventional planner (except possibly under strong crowding out), and the marginal benefits associated with increases in contributions induced by the incentive are correspondingly smaller. But the conventional planner does not take account of the direct effect of the subsidy on values, which in the case of crowding in will reduce the marginal costs of the subsidy (the right hand side of (16)). So one cannot say which planner will adopt the greater incentive.

The effect of non-separability on the level of incentive adopted by the conventional planner is also ambiguous. Since the marginal net benefit of the subsidy (the left hand side of (19)) can increase or decrease with variations in δ, we cannot in general determine whether the optimal incentive is larger or smaller than the incentive assuming separability. But to demonstrate that either effect is possible it will be sufficient to study a plausible example assuming a specific subsidy cost function, namely $c'(s) = \tau s$ and a linear benefit function, ϕ'' =0. In this case we can solve (19) for s, giving

(20) $$s^*(\delta) = \frac{(1+\delta)(n\phi' - \phi' - \underline{v})}{g''\tau + (1+\delta)^2}$$

We find the following expression for the condition for the over-use of incentives, $s^s - s^*(\delta) > 0$ (see Appendix).

(21) $$s^s - s^*(\delta) = \kappa(\frac{(1+\delta)}{g''} - \tau)\delta \qquad \text{for some postive } \kappa$$

Because κ is positive, we can determine which subsidy is greater depending on the parameters, δ and τ. This gives us proposition 3.

Proposition 3. Over-use or under-use of incentives by the naive conventional planner under crowding out. *The naive conventional planner may under-use incentives when either* *crowding in or crowding out hold; and also may overuse incentives for crowding in and crowding out.*

We demonstrate the proposition by an example. *If* $c'(s) = \tau s$ *and* $\phi'' = 0$, *then*

$$s^s > s^*(\delta) \text{ if and only if } (\frac{(1+\delta)}{g''} - \tau)\delta > 0$$

Figure 1 illustrates proposition 3. The shaded areas indicate the combinations for which we have the counter-intuitive result that if crowding out obtains, the naive planner will under-use incentives and conversely.

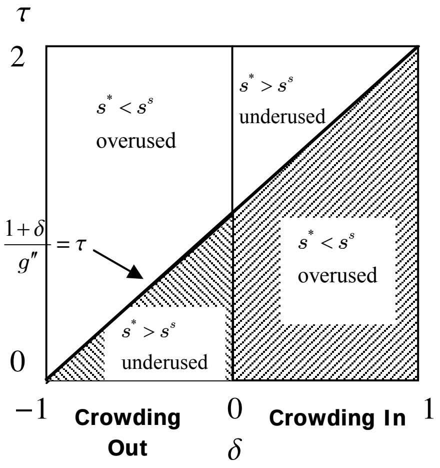

Figure 1. When are explicit incentives over-used by the naive planner? In shaded regions the subsidy is relatively effective ($(1+\delta)/g'' > \tau$), and the optimal choice of conventional social planner is different from the revealed preference social planner. The figure depicts the case where $g'' = 1$. Larger groups or less concave or more convex returns to the project would enlarge the shaded area.

The economic intuition behind these results is readily explained. The optimal subsidy as given by (16) for the revealed preference planner and (19) for the conventional planner is that which equates the marginal rate of transformation of the subsidy into contributions to the marginal rate of substitution between subsidies and contributions in the planner's objective function or

(22) $$\mathrm{MRT}=\frac{da^{*}}{ds}=\frac{-\partial\omega/\partial s}{\partial\omega/\partial a}=\mathrm{MRS}$$

For the revealed preference planner this gives

(23) $$\frac{1+\delta}{g''-n\phi''}=\frac{c'-a^{*}\delta}{(n-1)\phi'-s}$$

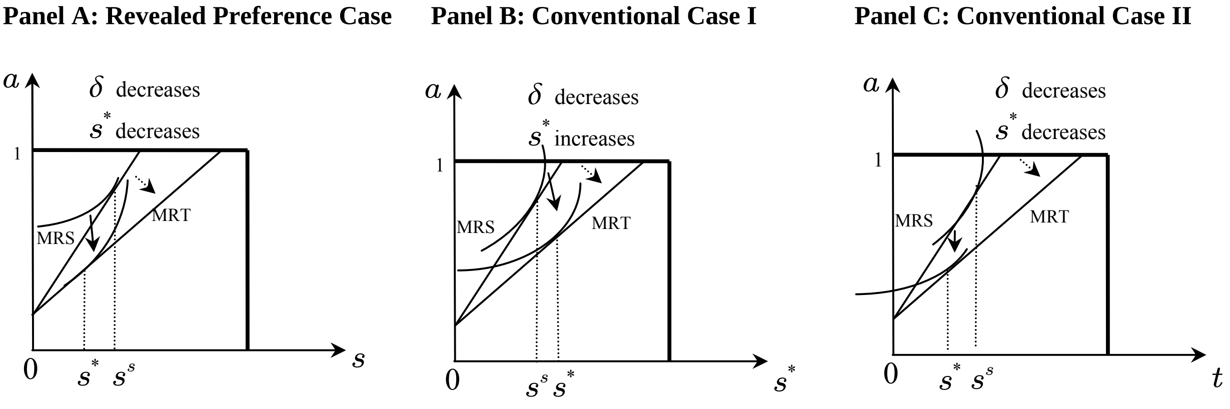

Figure 2. Effect of crowding out on optimal incentives. Solid arrows show changes in MRS and dotted arrows show changes in MRT.

As figure 2 panel A shows, a decrease in δ reduces the MRT and increases $-\partial\omega/\partial s$ without affecting $\partial\omega/\partial a$ thus raising the MRS and resulting in an unambiguous negative effect on the optimal subsidy. By contrast, for the conventional planner the MRS = MRT condition is

(24) $$\frac{1+\delta}{g''-n\phi''}=\frac{c'}{(n-1)\phi'-s-\underline{v}-\delta s}$$

so crowding out (as shown in figure 2 panel B) reduces the MRT as before but it increases $\partial\omega/\partial a$ while not affecting $\partial\omega/\partial s$, thus lowering the MRS. As lower values of δ 'flatten' the marginal rate of transformation while 'steepening' the planner's indifference loci, the effect on s^{*} is ambiguous, and depends of which effect – lowering the MRT (by $1/(g''-n\phi'')$) or lowering $\partial\omega/\partial a$ (by $\tau s/((n-1)\phi'-s-v-\delta s)^{2}$) – predominates. Proposition 3 gives the conditions for which effect will predominate.

5. *Conclusion: Public economics in light of behavioral economics*

Incentives work. This is particularly true of positive incentives to engage in activities for which there is little or no pre-existing motivation or ethical obligation, and for negative incentives that avoid conveying unfavorable information about the type or intentions of those with whom the individual is interacting. In some experiments, the magnitude of 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 (2005)), consistent with the separability assumption.

But the experimental evidence also suggests that the socially beneficial effects of public-spirited motives may be either enhanced or diminished by policy interventions that are designed by a naive social planner to more closely align self-regarding incentives with social objectives. We have shown that the naive social planner may over-use or under-use explicit incentives. If the planners problem is compliance with a target a higher level of incentive use is optimal if crowding out obtains (by comparison with the separable case) because crowding out makes the incentive less effective so that to attain the target, more incentive is needed. Correspondingly, when values and incentives are complements, a lower level of incentive is sufficient to induce the behavior ensuring compliance, so the required incentives are lower. By contrast, if the problem facing the planner is to maximize citizens' utility including their values, then over-use of incentives by the planner occurs when crowding out obtains, leading to policies that are less effective than anticipated, or (in the case of strong crowding out) may even be counter productive in that their effects are opposite of those intended. But for conventional planner (who takes account of the effect of values on behavior but does not include them in his own objective function) over-use of incentives may occur when values and incentives are either complements or substitutes, and the same is true of over-use.

One may conclude, then, that while explicit incentives do a tolerably good job in many situations, in others performance would be improved if mechanism design took account of the effects of incentives on values. Social preferences are a variable resource for the policy maker, one that may be either empowered or diminished by legislation and public policy.

This is the foundation of Hirschman's suggestion (quoted at the outset) that, counter to conventional economic logic, prohibitions may be superior to incentives of the type modeled here, even when the expected material cost of anti-social behavior is identical under the two regimes. The reason is that by explicitly proclaiming a behavior as anti social a prohibition may be complementary with individual's values, affirming a citizen's moral predisposition to not behave anti-socially rather than crowding out moral sentiments as may be the case of conventional incentives. Experimental evidence (Galbiati and Vertova, 2008) is consistent with this commonplace wisdom of legal theory (Kahan 1997).

Taking account of social preferences in mechanism design may be especially important in heterogeneous populations of both self-regarding and civic-minded individuals. In this case some mechanisms provide incentives that induce even the civic-minded to act as if they were selfish. Examples include anonymous competitive markets with parametric prices and public goods environments without opportunities for peer monitoring and sanctioning of non-contributors (Sobel, 2007, Fischbacher, Fong, and Fehr, 2005). Other mechanisms, such as the public goods game with peer punishment, may induce the self-interested to act as if they were civic-minded (Fehr and Gaechter, 2000).

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 critical information is non-verifiable so contracts are incomplete and the reach of governmental fiat is limited. The reason is that in these cases 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 seen, conventional 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 designed to more closely align self-regarding preferences and public objectives 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

Citation	*Subject pool*	*Game*	*Result*	*Comment*
Bohnet and Baytelman (2007)	Senior executives in U.S.	Trust Game: one shot, repeated, without and with punishment, communication ("institutions")	"Institutions" 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"
Bohnet, Frey, and Huck (200)	U.S. students	Contract enforcement	Compliance is non-monotonic in degree of enforcement	"Monetary" preferences crowd out "Honest" preferences where enforcement is moderately strong
Cardenas, Stranlund, and Willis (2000)	Colombian rural poor	Common pool resource with externally imposed fines	Fines induce more self-interested behavior and pool over-exploitation	Fine induced a shift from moral to self interested frame
Carpenter, Bowles, and Gintis (2007)	U.S. students	Public goods with peer punishment	Peer punishment induced defectors to contribute more, even when defection remained a best response	Peer 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"
Falk and Kosfeld (2005)	Swiss students	Trust Game	Principals who impose a minimum return rate on trustees receive less than trusting Ps	Imposed minimum understood by Subjects as a sign of distrust by Principals

Table 1 continued, next page

Table 1. Explicit Incentives and Social Preferences, continued

Citation	*Subject pool*	*Game*	*Result*	*Comment*
Fehr and Gaechter (2002)	Swiss students	Gift Exchange	Explicit incentives reduce effort (especially if negative), redistribute surplus to principal.	Framing and inequality aversion Incentives eliminate the positive effects of generosity
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
Fehr, Gachter, and Kirchsteiger (1997)	Swiss students	Gift Exchange (effort non- contractible)	Monitoring and fines reduced effort	
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 agents may [explain] why many contracts are ...left incomplete'
Fehr and List (2004)	Costa Rican CEO's & students	Trust Game with optional punishment	Not using the punishment option when it is available results in high performance	Key to performance: "the psychological message .. conveyed by incentives – whether ... kind or hostile..."
Fischbacher, Fong, and Fehr (2005)	Swiss students	Bilateral "Bargaining" vs "Market" versions of Ultimatum Game	Competition among respondents lowered offers, reduced rejections	Competition made punishment of 'unfair' offers less certain

Table 1 continued, next page

Table 1. Explicit Incentives and Social Preferences , continued

Citation	*Subject pool*	*Game*	*Result*	*Comment*
Frohlich and Oppenheim (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
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
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”
Galbiati and Vertova (2007)	Italian students	Public goods game with rewards and penalties	An externally announced contribution norm raises contributions independently of self-regarding incentives.	Contributions respond to socially determined ‘obligations’
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)
Gneezy and Rustichini (2000b)	Israeli students	Payment for soliciting contributions to social causes	Payment may reduce the performance of the solicitors	

Table 1 continued, next page

Table 1. Explicit Incentives and Social Preferences , continued

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
Houser, Xiao, McCabe, et al. (2007)	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."
Henrich, Boyd, Bowles, et al. (2005)	hunters, gatherers, herders, farmers in15 societies	Ultimatum Game	Offers and rejection of low offers were greater in more market-integrated societies	endogenous preferences: markets may have supported fair-mindedness towards strangers "*doux commerce*"? Hirschman (1977)
Hoffman, McCabe, Shachat et at. (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 (2005)	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"

Table 1 continued, next page

Table 1. Explicit Incentives and Social Preferences, continued

Citation	*Subject pool*	*Game*	*Result*	*Comment*
Rodriguez-Sickert, Guzman, and Cardenas (2007)	Rural Colombian adults	Common pool resource game with low and high fines for overexploitation	Low fines as effective as high fines	"Prescriptive effect" of the fine dominates the "guilt relief effect". Small fines crowd in unconditional cooperation by relieving cooperators of the need to retaliate against defectors.
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
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
Upton (1974)	U.S blood donors	Paid donations or uncompensated	Highly motivated givers respond negatively to incentives	See Titmuss (1971), Bliss (1972), Arrow (1972)

Appendix

1. *Proof* of proposition 2 b) and c). We suppose that the social planner's maximization problem is well defined so that the second order condition is satisfied. Equation (16) defines implicitly the optimal subsidy s^* depending δ :

$$(25)\qquad \left[(n-1)\phi'(na^*(s^*))-s^*\right]\frac{1+\delta}{g''(a^*(s^*))-n\phi''(na^*(s^*))}+a^*(s^*)\delta-c'(s^*)=0$$

Using the implicit function theorem and (12) and rearranging give

$$(26)\qquad \frac{ds^*}{d\delta}=\frac{((n-1)\phi'-s^*)\frac{1}{g''-n\phi''}+a^*}{c''+\frac{1-\delta^2}{g''-n\phi''}-n(n-1)\phi''(\frac{1+\delta}{g''-n\phi''})^2+((n-1)\phi'-s^*)(g'''-n^2\phi''')\frac{(1+\delta)^2}{(g''-n\phi'')^3}}$$

We verify that the denominator in equation (26) is equal to the second order condition of social planner's maximization problem except the change of sign, thus has a positive sign. We have assumptions 1 and 2 ($(n-1)\phi'>\underline{v}\geq s^*$), and the condition for the asymptotic stability of the Nash equilibrium ($g''-n\phi''>0$ in equation (12)) imply that the numerator is positive and thus we find $ds^*/d\delta>0$ ■

2. Expression (21)

$$(27)\qquad \begin{aligned} &s^s-s^*(\delta)=\kappa(\frac{(1+\delta)}{g''}-\tau)\delta \qquad \text{for some postive } \kappa \\ &\text{where } \kappa=\frac{(n\phi'-\phi'-\underline{v})g''}{(g''\tau+1)(g''\tau+(1+\delta)^2)} \end{aligned}$$

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.

Bar-Gill, Oren and Chaim Fershtman. 2005. "The Limit of Public Policy: Endogenous Preferences." *Journal of Public Economic Theory*, Vol. 7 (5), 2005, pp.841-857.

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

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.

Camerer, Colin. 2003. *Behavioral Game Theory: Experimental Studies of Strategic Interaction*. Princeton: Princeton University 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. 2007. "Strong Reciprocity and Team Production."

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

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.

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*.

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

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

Fehr, Ernst, Alexander Klein, and Klaus M. Schmidt. 2007. "Fairness and Contract design." *Econometrica*, 75:1, pp. 121-54.

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

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 Simon Gaechter. 2000. "Cooperation and Punishment in Public Goods Experiments." *American Economic Review*, 90:4, pp. 980-94.

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

Fehr, Ernst, Simon Gaechter, and Georg Kirchsteiger. 1997. "Reciprocity as a Contract Enforcement Device: Experimental Evidence." *Econometrica*, 65:4, pp. 833-60.

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, Bruno S. 1994. "How Intrinsic Motivation Is Crowded Out and In." *Rationality and Society*, 6:3, pp. 334-52.

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. 2007. "Behavioral Effects of Obligations." *Bocconi University*.

Galbiati, Roberto and Pietro Vertova. 2008 (forthcoming). "Obligations and Cooperation in Public Goods Games. " *Games and Economic Behavior.*

Gintis, Herbert, Samuel Bowles, Robert Boyd, and Ernst Fehr, Eds. 2005. *Moral sentiments and Material Interests: The Foundations of Cooperation in Economic Life*. Cambridge, MIT Press.

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.

Henrich, Joe, Robert Boyd, Samuel Bowles, et al. 2005. "'Economic Man' in Cross-Cultural Perspective: Behavioral experiments in 15 small-scale societies." *Behavioral and Brain Sciences*, 28.

Hirschman, Albert O. 1985. "Against parsimony: three ways of complicating some categories of economic discourse." *Economics and Philosophy* 1(1): 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.

Houser, Daniel, Erte Xiao, Kevin McCabe, et al. 2007. "When Punishment Fails: Research on Sanctions, Intentions, and Non Cooperation." *Games and Economic Behavior*, in press.

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

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

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

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.

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

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

Rodriguez-Sickert, Carlos, Ricardo A. Guzman, and Juan Camilo Cardenas. 2007. “Institutions influence preferences: evidence from a common pool resource experiment”, *Journal of Economic Behavior and Organization*. (in press)

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.

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

Sobel, Joel. 2007. Do markets make people selfish? (University of California at San Diego)

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

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

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.