UNIVERSITÀ DEGLI STUDI DI SIENA

QUADERNI DEL DIPARTIMENTO
DI ECONOMIA POLITICA

Samuel Bowles
Sung-Ha Hwang

Social Preferences and Public Economics: Mechanism Design when Social Preferences Depend on Incentives

n. 530 – Marzo 2008

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 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 sophisticated planner cognizant of these non-additive effects would make either more or less use of explicit incentives, by comparison to a naive planner who assumes they are absent.

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

We would like to thank Margaret Alexander, Lopamudra Banerjee, Ernst Fehr, Duncan Foley, John Geanakoplos, Suresh Naidu, Seung-Yun Oh, Carlos Rodriguez-Sickert, Sandra Polania Reyes, John Roemer, Bob Rowthorn, Paul Seabright, Rajiv Sethi, Joaquim Silvestre, Peter Skott, Joel Sobel, E. Somanathan, Tim Taylor, Elisabeth Wood, Giulio Zanella and members of the Yale Law School Legal Theory Seminar for their contributions to this research. Thanks also to the Behavioral Sciences Program of the Santa Fe Institute, the U.S. National Science Foundation, the European Science Fondation and the University of Siena for financial support of this project.

Samuel Bowles, Santa Fe Institute and Department of Economics, University of Siena
Sung-Ha Hwang, Depts. of Economics and Mathematics, 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 is 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 supposition is "false in fact": altruism, reciprocity, and what the classicals called civic virtues 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), generalized obedience to law (Kahan, 1997), and other areas critical to public economics.

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. What Adam Smith termed "the moral sentiments" played a central role in their thinking. But they assumed that ethical motivations would be unaffected by incentive-based policies designed to recruit self-interest to public ends. Along with civic virtues, explicit incentives and constraints could thus contribute additively to good government. According to this view, taxes or subsidies affect individual utility and hence behavior only indirectly, that is 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. As a result the behavioral effects of moral sentiments and the material interests are separable, the effects of each being independent of the levels of the other. But when separability does not hold, the two kinds of motivations may be either complements -- social preferences being heightened by incentives appealing to self interest -- or substitutes, when explicit incentives are said to crowd out social preferences.

A consequence of the classicals' implicit 'separability assumption' is that they failed to take account of how harnessing self -interest to the public good might either compromise or enhance civic virtues. While in contemporary economic theory separability is not explicitly assumed and could be abandoned, modern public economics, mechanism design and related fields continue the classicals' practice. However a great many experiments and observations in natural settings suggest that social preferences are often important influences on behavior, and that the salience of these preferences varies with the kinds of explicit incentives that are implemented.

If the separability assumption is false, policies designed on its basis will generally be non-optimal, and explicit incentives will be over-used or under-used. Over-use of explicit incentives when crowding out is the case 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), Bowles (1998), Cooter (1998), Ostrom (2000), and Bar-Gill and Fershtman (2005).

The economic intuition underlying Titmuss' and Hirschman's concerns is that because crowding out reduces the effectiveness of explicit incentives, they would be used less by a sophisticated social planner cognizant of the crowding out problem, by comparison to a naive planner, namely, one who assumes that economic and moral motives are separable. If crowding out is so strong that the incentive has an effect the opposite of its intent, this is of course the case. But the effect of crowding out need not be literally counterproductive in this sense and where the effectiveness of incentives is blunted but not reversed, the implications for the optimal use of incentives are 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. To do this we develop a model of optimal explicit incentives in the presence of both crowding in and crowding out, and use the model to identify cases in which crowding out entails greater or lesser use of incentives.

To analyze these cases we will ask what incentives would be adopted by a social planner who wishes to maximize the aggregate utility of citizens. (By "incentives" without adjective we mean those appealing to conventional self-regarding preferences.) 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.

In the next section we survey the empirical literature on non-separability. We then introduce a model of public incentives when individuals with social preferences may contribute to a public good, using this model to clarify the separability assumption and how it may be violated. In section 4 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 holds. In section 5 we study optimal incentives for the sophisticated planner who maximizes total social welfare, including the values of the citizens

both as components of their utility and influences on their behavior. We find that in a public goods setting, as in the compliance case, the sophisticated planner may make more use of incentives than the naive planner when crowding out is the case. The economic intuition behind this surprising result is evident in the compliance case. In cases where the marginal social benefits of the public good rise sharply as the shortfall from its socially optimal level increases (a limiting case of which is the target compliance problem), the fact that crowding out makes the incentive less effective requires its greater use. Where decreasing marginal returns to the public good are modest or absent the sophisticated planner will make lesser use of incentives when crowding out holds, as expected. In section 6 we consider the implications of non-separability for public economics.

2. When separability fails; evidence and explanations

The underlying social and psychological mechanisms accounting for non-separability include the following. (Bowles(2008) and Frey and Jegen (2001) survey the experimental and other evidence.)

First, 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 ; Tversky and Kahneman, 1981).

Second, the incentives adopted by a principal unavoidably provide information about the principal's preferences as well as the nature of the task to be done and his beliefs about the trustworthiness of the agent or other aspects of the agent's likely behavior (Benabou and Tirole, 2003 ; Seabright, 2004). The use of explicit incentives may thus 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). The presence of incentives may also reduce the value of generous or civic minded acts as a signal of one's moral character (Benabou and Tirole, 2006).

Third, 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 (Deci, Koestner, and Ryan, 1999; Cameron, Banko, and Pierce, 2001; Frey, 1994). The experiment of Mellstrom and Johannesson, (2008) suggests that Titmuss may have been right about this for the case of women potential blood donors (but not men).

Fourth, the incentives adopted by a principal influence the process by which agents invest in identities and update their preferences and may bias it in a self-interested direction (Ben-Porath, 1980; Bohnet, Frey, and Huck, 2001; Bowles, 1998; Falkinger et. al., 2000; Gaechter, Kessler, and Konigstein, 2007; Bar-Gill and Fershtman, 2005).

Fifth, 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; Gaechter and Falk, 2002; Rodriguez-Sickert, Guzman, and Cardenas, 2007).

3. Moral sentiments and material interests as complements or substitutes

We abstract from these 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 non-negative, increasing and convex in its argument. The output of the project depends on each member's contribution, $\phi(a^1, a^2, ..., a^n)$ and explicit

incentives take the form of a subsidy $s \geq 0$ proportional to the amount contributed. Implementing the subsidy entails administrative, 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 their contribution level. We suppose that payment of the taxes supporting the subsidy has no effect on citizens' behavior and can be ignored. The net social cost of the subsidy is thus just what we call its administrative cost namely, $c(s)$.

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 and allow its representation in a single parameter we abstract from individual differences in the effects of incentives on values and give the values function an explicit form

(1) $$v = a^i(\underline{v} + \lambda s)$$

so the marginal effect of i's contributing on i's values is $v_{a^i} = \underline{v} + \lambda s$. The classical separability assumption maintains that the level of explicit material incentives does not influence the marginal value utility of contributing: that is $\lambda = 0$. We do not consider the case of taxes (i.e. $s < 0$) because motivational crowding is not symmetrical: in experiments, both bonuses and fines crowd out social preferences (thought typically in different degree) so one cannot reverse the crowding effect by adopting taxes rather than subsidies.

Not all of the complex psychological mechanisms accounting for non-separability are captured by this simple formulation; for example it precludes plausible cases in which simply the presence of the incentive has a substantial effect on values even if the incentive is arbitrarily small (Gneezy and Rustichini, 2000b) or where the effect of incentives on values depends on the actions or values of others. For example in the employer employee gift exchange experiment of Irlenbusch and Sliwka (2005) subjects' effort responses to variations in piece rates closely

approximated those predicted on the basis of simple payoff maximization, but effort levels were higher in the complete absence of piece rates, the apparent framing effect of which negatively affected motivation, equivalent to a shift downwards of $\underline{v}$ in our equation (1). However our formulation illustrates the fundamental problem when values and incentives are either complements or substitutes and provides a tractable way to study the implications for mechanism design.

Using (1) individual i's utility is

$$u^i = \phi\left(a^1, a^2, .., a^n\right) + sa^i - g(a^i) + v(a^i, s) \tag{2}$$

Varying a^i to maximize u^i for given values of s and the others' contributions, the individual's best response a^i is given by

$$g'(a^i) = \phi_{a^i} + s + \underline{v} + \lambda s \tag{3}$$

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 contribution. To rule out corner solutions we assume throughout that $g'(1)$ and $g'(0)$ are (respectively) sufficiently large and sufficiently small so that the value of a satisfying the citizen's best responses lies in the unit interval.

From (3) the effect of the subsidy on the individual's contribution (given the contributions of others) is then

$$\frac{\partial a^i}{\partial s} = \frac{1+\lambda}{g'' - \phi_{a^i a^i}} \tag{4}$$

where the denominator is positive by the second order condition of the individual's optimization problem (in the case of a convex benefit function for the public project, requiring that the marginal costs of contributing be rising faster than the marginal private material benefits).

Where the separability condition does not hold, we have either crowding in ($\lambda > 0$) or crowding out ($\lambda < 0$). 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 $\lambda < -1$, which we term strong crowding out, the incentive reduces contributions. Strong crowding out is evident the Haifa day care case and other experiments surveyed in Bowles (2008) in which incentives had the opposite of the intended effect. But it is clear from equation (4) that a positive response by subjects to explicit incentives does not indicate that crowding out is absent; it indicates only that $\lambda > -1$.

We can now clarify the distinction between the naive and the sophisticated planner. The subsidy adopted by the naive planner who assumes separability is denoted, s^{N} and this subsidy is obviously equal to the sophisticated planner's optimal subsidy $s^{*}(\lambda)$ in the case that the separability assumption is true, so $s^{\mathrm{N}} = s^{*}(0)$. Then we say that incentives are under-used if $s^{*} > s^{\mathrm{N}}$ and conversely.

Because we wish to model the under-provision of a public good when only private incentives are in force, and the possible implementation of a superior outcome through a publically implemented incentive, we make the following assumptions

1. Values alone are insufficient to internalize the external benefits of contributing to the public good: in the absence of a subsidy, the marginal benefits that one's contributions confer on others in the community exceed the marginal value utility of contributing, or $(n-1)\phi_{a^i} > \underline{v}$

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

4. Ensuing compliance

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) $$\phi(a^1, a^2, ..., a^n) = \sum_i \phi^i a^i$$

where ϕ^i is a constant that may differ among individuals as the public benefits of an individual having first aid knowledge differ. Then individual i's utility is

(6) $$u^i = \sum_j \phi^j a^j + s a^i - g(a^i) + a^i \underline{v} + a^i \lambda s$$

Therefore the individual's best response is given by

(7) $$g'(a^i) = \phi^i + s + \underline{v} + \lambda s$$

To identify the marginal individual (assumed to be unique) who must contribute $\bar{a}$ in order to secure the compliance target of the planner we reorder the index such that $\phi^i \leq \phi^j$ for $i < j$. The marginal individual is then $\bar{i}$ where $\bar{i}$ is the smallest number, i, satisfying $i > n(1-p)$. The case of interest is that in which the critical individual's values and own benefits from contributing are insufficient to motivate his attaining the target in the absence of the subsidy (that is $g'(\bar{a}) > \phi^{\bar{i}} + \underline{v}$). Then the social planner will choose $s^*(\lambda) = 0$ if $\lambda \leq -1$, abandoning the target as unattainable by use of the subsidy, and otherwise select the subsidy satisfying

(8) $$g'(\bar{a}) \le \phi^{\bar{i}} + s^*(\lambda) + \underline{v} + \lambda s^*(\lambda)$$

Since providing the subsidy is costly, if it is used at all the social planner will choose the minimum $s^*(\lambda)$ satisfying (8).

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

The naive planner believes that $\lambda = 0$ and hence adopts $s^{\mathrm{N}} = g'(\bar{a}) - (\phi^{\bar{i}} + \underline{v})$ as his preferred subsidy. From (9) we have

$$s^{\mathrm{N}} < s^*(\lambda) \quad \text{if and only if} \quad -1<\lambda<0$$

In case of crowding out (in), the sophisticated planner uses the incentive more (less) than the naive planner.

5. *Optimal incentives for the provision of a public good*

We turn now to the problem of the planner who 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 to a public good may 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) $$\phi(a^1, a^2, ..., a^n) = \phi\left(\sum\nolimits_j a^j\right)$$

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 derive the individual citizen's best response as in the case of equation (3) and solve for all of the contribution levels, a^i 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

the individual's Nash equilibrium contribution (denoted as a^*, suppressing the individual subscript) satisfies the following condition:

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

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

(12) $$\frac{\partial a^*}{\partial s} = \frac{1+\lambda}{g'' - n\phi''}$$

where the derivatives of g'' and ϕ'' are evaluated at a^* and na^* respectively and the asymptotic stability of the Nash equilibrium requires the denominator to be positive. Equation (12) differs from the partial effect of the subsidy on an individual's contribution (e.g. equation (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 (10) is concave and multiplied if it is convex, as expected. Like equation (4), equation (12) confirms 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 behavior of a single citizen in response to the planner's choice of s to maximize the social welfare function:

(13) $$\omega(s) = \phi(na^*(s,\lambda)) - g(a^*(s,\lambda)) + v(a^*(s,\lambda), s) - c(s)$$

The optimal incentive is given by

(14) $$s^*(\lambda) = \arg\max_s \omega(s)$$

so the planner chooses a subsidy satisfying

(15) $$\left[n\phi'(na^*(s,\lambda)) - g'(a^*(s,\lambda)) + \underline{v} + \lambda s\right]\frac{\partial a^*}{\partial s} + a^*(s,\lambda)\lambda - c'(s) = 0$$

The first term in the left hand expression is the net indirect effect of the change in contributions induced by variation in the subsidy, showing that planner takes account of the fact that for the individual the value benefits partially offset the material costs of contributing. 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), we find that the individual's marginal cost of contributing net of the marginal value benefits, namely, $g' - \underline{v} - \lambda s$ is just $\phi' + s$. Making this substitution in (15) and using (12) we see that the optimal subsidy is either zero or the positive value of s satisfying (16) so as to equate marginal benefits of the subsidy to its marginal costs:

$$(16) \qquad \underbrace{\left[(n-1)\phi' - s\right]\frac{(1+\lambda)}{g'' - n\phi''} + a^*\lambda}_{\text{marginal benefit}} \; - \; \underbrace{c'(s)}_{\text{marginal cost}} \; = 0$$

where we suppress the arguments of ϕ', g'', ϕ'', a^*. The second order condition is satisfied if the marginal benefit function is declining in s and the marginal cost function is constant or increasing, as is shown in figure 1.

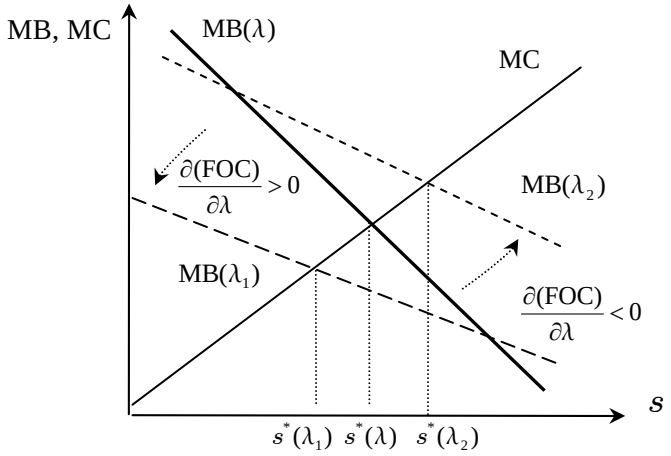

Figure 1. Effect of crowding out for the optimal incentives. The figure depicts equation (16), the determination of the planner's optimal incentive and the effect of a reduction in λ

To determine the effect of variations in λ on s^* we totally differentiate (16), the first order condition (FOC) with respect to s and λ and evaluate the result at s^*. Thus we have

(17) $$\frac{ds^*}{d\lambda} = -\frac{\partial(\text{FOC})/\partial\lambda}{\partial(\text{FOC})/\partial s}$$

From the second order condition of the planner's optimum problem we know that the denominator is negative, so the sign of the effect of non-separability on the optimal level of incentives is given by $\partial(\text{FOC})/\partial\lambda$, that is, whether crowding out – a decrease in in λ – shifts the marginal benefit function in figure 1 upwards or downward. In the case shown by the dashed line, crowding out shifts the marginal benefit function down and thus entails a lesser use of the incentive. This captures the economic logic of Titmuss' and Hirschman's critique of the use of incentives mentioned at the outset.

What drives this result is that a decline in λ reduces the effectiveness of the subsidy, which (as can be seen from equation (16)) reduces the marginal benefit of the subsidy. But closer inspection of equation (16) (see appendix) makes it clear that variations crowding out may have the opposite and less intuitive effect that crowding out induces greater use of the incentive. A reduction in λ reduces total contributions to the public good and if the marginal public benefits from contribution are diminishing in the total contributed ($\phi'' < 0$) then ϕ' will increase, possibly offsetting the reduced effectiveness of the incentive and shifting the marginal benefit function upwards, thus entailing a greater use of the subsidy.

Thus the sign of $\partial\text{FOC}/\partial\lambda$ cannot be determined in general, and crowding out may either increase or decrease the subsidy adopted by the sophisticated planner. To explore the counter-intuitive case in which crowding out results in greater use of the incentive, we adopted specific but plausible utility, cost, and public project functions and varied λ. Figure 2 shows results for a concave public goods benefits function, with two functions, one representing the planner's

marginal benefits of variations in the subsidy under crowding out, and the other under separability. Here crowding out results in an upward shift in the marginal benefit function resulting in greater use of the incentive and confirming the intuition in the previous paragraph. Panel B gives the citizen's best response function and the resulting levels of contribution under separability and crowding out. Panel C presents the optimal level of subsidy as a function of λ. Notice that as λ approaches -1 the optimal subsidy rises at an increasing rate, reaches a maximum and then declines to 0 (when λ = -1). The sharp decline is occasioned by the fact that as λ falls, the marginal benefits function becomes increasingly flat, so that shifts upwards or downwards in it produce increasingly large shifts in s^*. Panel D contrasts this case with one based on a linear public goods production function, in which, as anticipated s^* is monotonically increasing in λ (The simulation and analysis of equation (17) are presented in more technical detail in the appendix.)

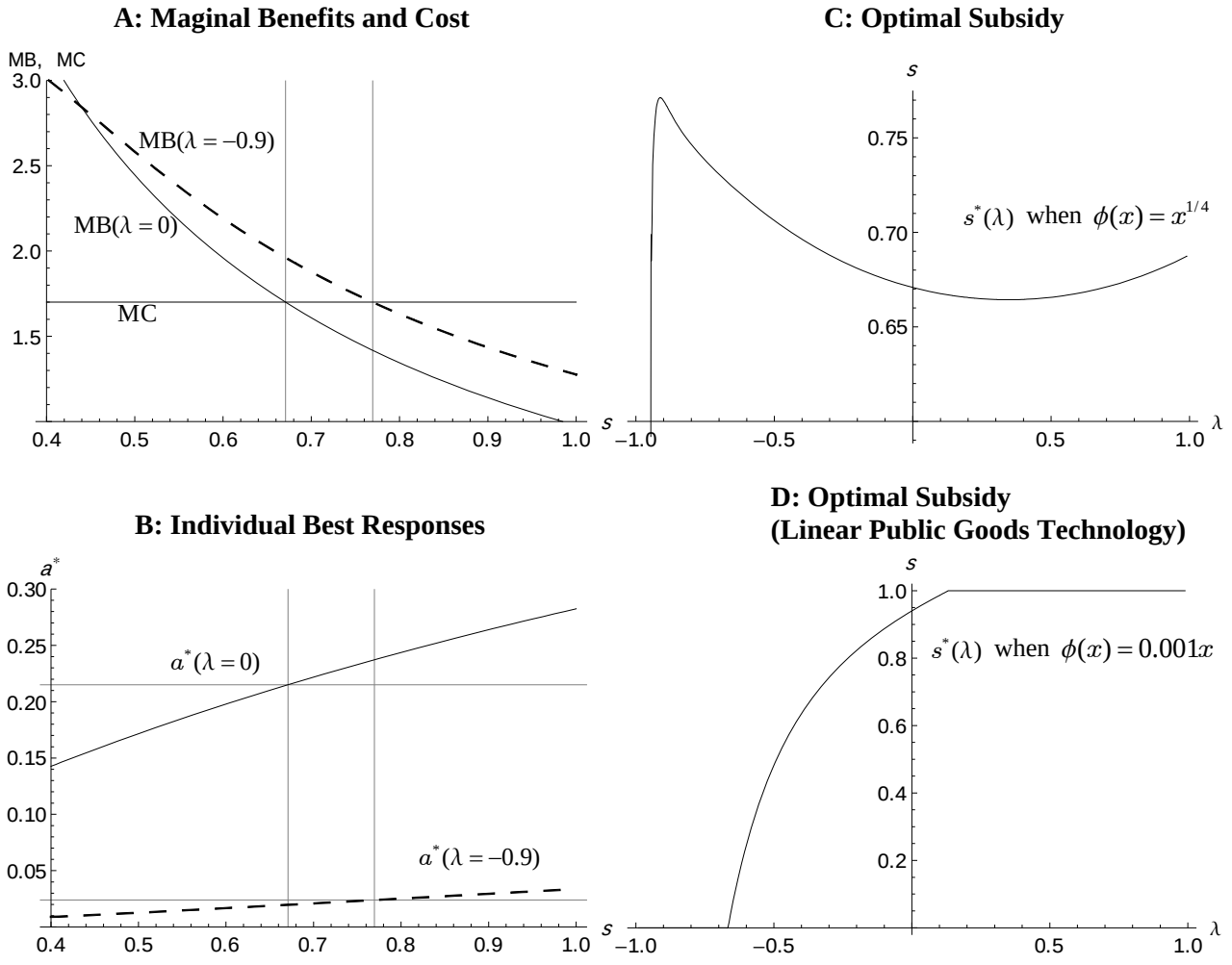

Figure 2. Under-use of incentives under crowding out. We use the following functions and parametric values for the simulation of $s^*(\lambda)$: denoting x as the total contribution of citizens, $\phi(x) = x^{1/4}$, $n = 10000$, $\underline{v} = 1$, $g(a) = 1.03a/(1-a)$, $c(s) = 1.7s$. Details of the computations are in the appendix. In Panel A, the marginal benefit and the marginal cost are presented; the downward solid line shows the marginal benefit when $\lambda = 0$ whereas the dashed line is the marginal benefit in case of $\lambda = -0.9$. The horizontal line in Panel A is the marginal cost. Panel B depicts the resulting contribution levels of citizens, namely the graph of $a^*(s)$ given values of λ. Grid lines in Panel A and Panel B show the optimal choices of s^* and the determination of a^* given the specified values of λ; $\lambda = 0$, $s^* = 0.671$, $a^* = 0.215$; $\lambda = -0.9$, $s^* = 0.769$, $a^* = 0.024$. Panel C shows the graph of $s^*(\lambda)$. Finally, Panel D shows $s^*(\lambda)$ when ϕ is linear: $\phi(x) = 0.001\,x$.

6. 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 the individual implementing the incentives. In some experiments, the magnitude of the response to variations in a given incentive structure closely approximates what one would expect based on conventional self-regarding preferences alone (for example, Anderhub, Gaechter, and Konigstein, 2002, Falkinger, et. al., 2000). 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 sophisticated planner may use an explicit incentive either more or less than a naive planner depending on the nature of the mechanism design problem. If the planner's problem is compliance with a target, a higher level of incentive use is optimal if crowding out holds (by comparison with the separable case, and as long as strong crowding out does not hold). The reason is that crowding out makes the incentive less effective, so that to attain the target, more incentive is needed. By contrast, if the problem is to maximize citizens' utility including their values, then the sophisticated will make either greater or lesser use of incentives by comparison to the naive planner when crowding out holds, leading to policies that are less effective than anticipated, or (in the case of strong crowding out) may even be counterproductive in that their effects are opposite of those intended. The sophisticated planner may make greater use of incentives when crowding out occurs if the benefit function exhibits strongly diminishing returns. The same result holds if the sophisticated planner (as above) takes account of the behavioral effects of non-separability, but does not include the citizens' values as a component of the social welfare function and maximizes the material net benefits of the project namely $\phi(na) - g(a) - c(s)$.

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 preferences. 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 marginal cost of anti-social behavior is identical under the two mechanisms. The reason is that by explicitly proclaiming a behavior as anti-social, a prohibition may be complementary with individual 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. The "obligation effect" on preference represents an upward shift in our value function induced by an increase in $\underline{v}$. Experimental evidence is consistent with this commonplace wisdom of legal theory (Kahan 1997). Roberto Galbiati and Pietro Vertova (2008, in press and 2008) show that subjects faced with fines for under-contributing and rewards for contributing more than a stated obligation respond positively to variations in the obligation despite the fact that the incentives to contribute are unaffected. The obligations effect works in part through the subjects' beliefs about what others will do, and in part through an independent effect of obligations on preferences.

Taking account of social preferences in mechanism design may be especially important in heterogeneous populations. Optimal design in these cases will typically involve more complex instruments than the uniform and linear subsidy we have studied. For example, if the sophisticated planner knew the crowding parameter λ (in equation(1)) of each citizen (assuming that these differed across individuals), differential subsidies could be devised to maximize the effect of the subsidy for a given cost, using equation (12). Such discrimination among citizens

based on their values requires information not generally available and in any case might violate liberal legal and ethical norms however and prove politically infeasible.

Of greater practical relevance are situations where citizens differ in $\underline{v}$, so that the population is made up of both self-regarding and civic-minded individuals as is suggested by experimental evidence. 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 as well as public goods environments without opportunities for peer monitoring and sanctioning of non-contributors (Sobel, 2007; Fischbacher, Fong, and Fehr, 2003). 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; Gaechter and Falk, 2002 ; Carpenter et al, 2008).

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 that contracts are incomplete and the reach of governmental fiat is limited. The reason is that in these cases as Kenneth 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 Richard Lipsey and Kevin 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 Taylor (1987) had earlier suggested that the Hobbesian state may produce Hobbesian man.

Appendix

1. Derivation of $a^*(s,\lambda)$

Given (10), the individual's best response a^i satisfies the following equation.

(A1) $$g'(a^i) = \phi'(\sum_j a^j) + s + \underline{v} + \lambda s \quad \text{for } i = 1,...,n$$

Equation (A1) defines implicitly the individual i's best response given others' contribution and by solving the n equations in (A1), we can find Nash equilibrium, $(a^{1*},....,a^{n*})$, for the public goods game among n citizens. Since we look for a symmetric Nash equilibrium, by setting $a^i = a^*$ for all i, we find the condition for a^* as in equation (11). Now equation (11) defines a^* implicitly in terms of s and λ and we denote this solution as $a^*(s,\lambda)$. To find the effect of the subsidy and crowding out on the individual's Nash contribution , we substitute $a^*(s,\lambda)$ for a^* in (11) and take the derivatives of this expression with respect to s and λ.

(A2) $$\frac{\partial a^*}{\partial s} = \frac{1+\lambda}{g'' - n\phi''}, \quad \frac{\partial a^*}{\partial \lambda} = \frac{s}{g'' - n\phi''}$$

2. Simulation of $s^*(\lambda)$

To construct $a^*(s,\lambda)$, we divide the domains of s, [0,1], and λ, [-1,1], into 100 subintervals, respectively (in total 100^2 rectangles), solve for the optimal choices of citizens given each value of s and λ at the endpoints of these subintervals, and construct an interpolation of these values. Using the resulting values of $a^*(s,\lambda)$ (shown for two values of λ in figure 2B), we find the optimal choices of social planners given λ (using 1000 subdivisions of interval [-1,1]) and interpolate s^* to obtain the function $s^*(\lambda)$ (shown in figure 2C for the concave benefits function and in 2D for the linear benefits function).

3. First order condition for social planner's optimization problem

$$\frac{\partial \text{FOC}}{\partial \lambda} = \underbrace{((n-1)\phi' - s)\frac{1}{g'' - n\phi''}}_{\text{I}} + \underbrace{n(n-1)\phi''\frac{s(1+\lambda)}{(g'' - n\phi'')^2}}_{\text{II}} + \underbrace{\frac{s\lambda}{g'' - n\phi''} + a^*}_{\text{III}} - \underbrace{((n-1)\phi' - s)(1+\lambda)s\frac{(g''' - n^2\phi''')}{\left(g'' - n\phi''\right)^3}}_{\text{IV}} \quad \text{(A3)}$$

where we suppress the arguments of ϕ', g'', ϕ'', a^*

The first term (I) represents variations in the effectiveness of the incentive induced by the variation in λ multiplied by the marginal social benefits of contributions. This term must be non-negative by assumptions 1 and 2. The second term (II) represents the variation in these marginal benefits associated with the change in the level of contributions induced by the variation in λ. This will be negative if the benefit function is concave: a decrease in λ, for example, induces lesser contributions, which in this case raise the marginal social benefits of contribution. The third term (III) is the effect of variations in λ on the direct effect of the incentive on values, composed itself of a direct effect of variations in λ (that is, a^*) and an indirect effect via the effect of variations in λ on the level of contribution ($\lambda\, s/(g'' - n\phi'')$). This term is non-negative for $\lambda = 0$ but in general may have either sign. The last term (IV) represents the effect of the non-linearity of the citizens' best response function $a^*(s)$. The term, $g''' - n^2\phi'''$, will be zero if both g''' and ϕ'' are zero (and as a result $a^*(s)$ is linear), while taking a positive sign if $a^*(s)$ is concave. As expected, concavity of the individual's best response function works in the same direction as concavity of the benefits function.

The magnitudes of these four terms vary with λ as indicated in figure A1 panel A. The sum of these terms is shown in panel B, positive values indicating an upward shift in the marginal benefit function induced by an increase in λ, and a zero value occuring at the values of λ for which (from figure 2) s^* is at a maximum ($\lambda = -0.912$) and at a local minimum

($\lambda = 0.345$). At the levels of s^* implied by these values of λ variations in λ rotate the marginal benefit function around the point at which it intersects the marginal cost function leaving s^* unchanged.

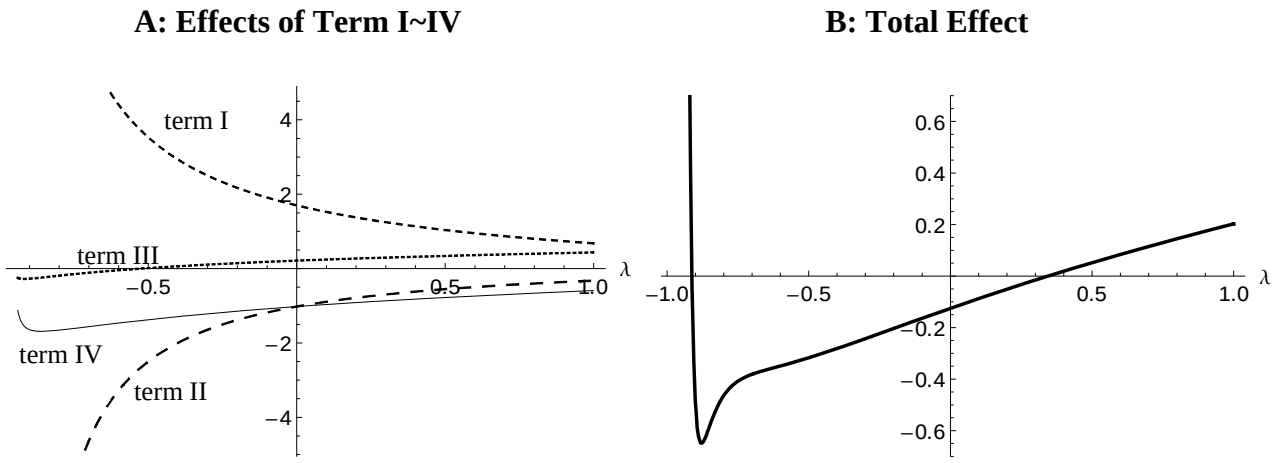

Figure A1. Effects of term I~IV and total effect. The functions and parameters used: $\phi(x) = x^{1/4}$, $n = 10000$, $\underline{v} = 1$, $g(a) = 1.03a/(1-a)$, $c(s) = 1.7s$.

4. Maximum optimal subsidy at high rates of crowding out.

The explanation in the text of the determination of s^* for values of λ approaching strong crowding out is illustrated in figure A2 showing the changes in the marginal benefit functions depending various values of λ, and hence the determination of optimal incentives, s^*. Notice that as λ moves from 0 to -.912, s^* increases and then as λ falls still further to -.94 it declines.

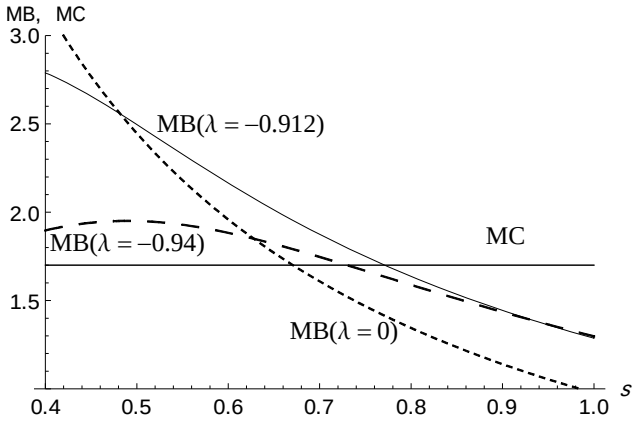

Figure A2. The functions and parameters used: $\phi(x) = x^{1/4}$, $n = 10000$, $\underline{v} = 1$, $g(a) = 1.03a/(1-a)$, $c(s) = 1.7s$. The optimal incentives are as follows: $\lambda = 0$, $s^* = 0.671$; $\lambda = -0.912$, $s^* = 0.771$; $\lambda = -0.94$, $s^* = 0.730$

Works cited

Anderhub, V., Gaechter S., Konigstein M., 2002. Efficient Contracting and Fair Play in a Simple Principal Agent Experiment. Experimental Economics 5 (1), 5-27.

Andreoni, J., Erand B., Feinstein J., 1998. Tax Compliance. Journal of Economic Literature 36 (2), 818-60.

Arrow, K. J., 1971. Political and Economic Evaluation of Social Effects and Externalities. In: M. D. Intriligator (Ed.). Frontiers of Quantitative Economics. Amsterdam: North Holland, 3-23.

Bar-Gill, O., Fershtman, C., 2005. The Limit of Public Policy: Endogenous Preferences. Journal of Public Economic Theory 7 (5), 841-857.

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

Benabou, R., Tirole, J., 2006. Incentives and Prosocial Behavior. American Economic Review 96 (5), 1652-1678

Ben-Porath Y., 1980. The F-Connection: Families, Friends, and Firms and the Organization of Exchange. Population and Development Review 6(1), 1-30.

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

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

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

Bowles, S., 2008. Policies designed for self-interested citizens may undermine “the moral sentiments”:evidence from economic experiments. Science, in press.

Camerer, C., 2003. Behavioral Game Theory: Experimental Studies of Strategic Interaction. Princeton: Princeton University Press.

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

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

Carpenter, J., Bowles, S., Gintis, H., 2008. Strong Reciprocity and Team Production. Working Paper, Santa Fe Institute

Cooter, R., 1998. Expressive Law and Economics. Journal of Legal Studies 27, 585-608.

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

Falk, A., Kosfeld, M., 2006. The Hidden Costs of Control. American Economic Review 96 (5), 1611-30.

Falkinger, J., Fehr, E., Gaecter, S., Winter-Ebmer, R., 2000. A simple mechanism for the efficient provision of public goods. American Economic Review, 90 (1), 247-64

Fehr, E., Gaechter, S., 2000. Cooperation and Punishment in Public Goods Experiments. American Economic Review 90 (4), 980-94.

Fehr, E., Klein, A., Schmidt, K. M., 2007. Fairness and Contract design. Econometrica 75 (1), 121-54.

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

Fehr, E., Rockenbach, B., 2003. Detrimental effects of sanctions on human altruism. Nature 422 (13) March, 137-40.

Fischbacher, U., Fong, C., Fehr, E., 2003. Fairness, errors, and the power of competition. Zurich, IERE Working Paper No 133. http://www.iew.uzh.ch/wp/iewwp133.pdf

Fong, C., Bowles, S., Gintis, H., 2005. Strong reciprocity and the welfare state. In: Serge-Christophe Kolm and Jean Mercier Ythier (Eds.). Handbook of Giving, Reciprocity, and Altrusim. Amsterdam: Elsevier, 1439-1464.

Frey, B. S., 1994. How Intrinsic Motivation Is Crowded Out and In. Rationality and Society 6 (3), 334-52.

Frey, B. S., 1997. A Constitution for Knaves Crowds Out Civic Virtues. Economic Journal 107 (443), 1043-53.

Frey, B. S., Jegen, R., 2001. Motivation Crowding Theory. Journal of Economic Surveys 15 (5), 589-611

Frohlich, N., Oppenheimer, J. A., 1995. The Incompatibility of Incentive Compatible Devices and Ethical Behavior: Some Experimental Results and Insights. Public Choice Studies, 25, 24-51.

Gaechter, S., Falk, A., 2002. Reputation or Reciprocity? Consequences for the Labour Relation. Scandinavian Journal of Economics, 104 (1), 1 - 26.

Gaechter, S., Kessler, E., Konigstein, M., 2007. Performance Incentives and the Dynamics of Voluntary Cooperation. University of Nottingham, School of Economics. http://www.eea-esem.com/EEA-ESEM/2006/Prog/viewpaper.asp?pid=2640

Galbiati, R., Vertova, P., 2008. Obligations and Cooperative Behaviour in Public Good Games. Games and Economic Behavior, doi: 10.1016/j.geb.2007.09.004, in press.

Galbiati, R., Vertova, P., 2008. Behavioral Effects of Obiligations. European University Institute: Firenze

Gintis, H., Bowles, S., Boyd, R., Fehr, E. (Eds), 2005. Moral sentiments and Material Interests: The Foundations of Cooperation in Economic Life. Cambridge: MIT Press.

Gneezy, U., Rustichini, A., 2000a. A Fine is a Price. Journal of Legal Studies 29 (1), 1-17.

Gneezy, U., Rustichini, A., 2000b. Pay enough or don't pay at all. Quarterly Journal of Economics 115 (2), 791-810.

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

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

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

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

Kahan, D. M., 1997. Social Influence, Social Meaning, and Deterrence. Virginia Law Review 83 (2), 349- 95.

Kreps, D. M., 1997. Intrinsic motivation and extrinsic incentives. American Economic Review 87 (2), 359-64.

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

Lazear, E., 2000. Performance Pay and Productivity. American Economic Review 90 (5), 1346 - 61.

Lipsey, R., Lancaster, K., 1956-1957. The General Theory of the Second Best. Review of Economic Studies 24 (1), 11-32.

Mellstrom, C., Johannesson, M., 2008. "Crowding out in blood donation: Was Titmus right?" American Economic Review, in press.

Ostrom, E., 2000. Crowding out Citizenship. Scandinavian Political Studies 23(1), 3-16.

Pommerehne, W.W., Weck-Hannemann, H., 1996. Tax rates, tax administration and income tax evasion in Switzerland. Public Choice 88 (1-13), 161-70.

Rodriguez-Sickert, C., Guzman, R. A., Cardenas, J. C., 2007. Institutions influence preferences: evidence from a common pool resource experiment, Journal of Economic Behavior and Organization. doi:10.1016/j.jebo.2007.06.004

Seabright, P., 2004. Continuous Preferences Cab Cause Discontinuous Choices: An Application to the Impact of Incentives on Altruism. CEPR Discussion paper no. 4322 London. http://www.cepr.org/pubs/dps/DP4322.asp

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

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

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

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

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