JOURNAL OF PUBLIC ECONOMIC THEORY
Department of Economics 415 Calhoun Hall Vanderbilt University Nashville, TN 37235 USA (615) 343-8495 (fax)

Submission Number:JPET-11-00279

Optimal incentives with state-dependent preferences

Sung-ha Hwang
Sogang University

Samuel Bowles
Santa Fe Institute

Abstract

In both experimental and natural settings incentives sometimes under-perform, generating smaller effects on the targeted behaviors than would be predicted for entirely self-regarding agents. A parsimonious explanation is that incentives that appeal to self-regarding economic motives may crowd out non-economic motives such as altruism, reciprocity, intrinsic motivation, ethical values and other social preferences, leading to disappointing and sometimes even counter-productive incentive effects. We present evidence from behavioral experiments that crowding may take two forms: categorical (the effect on preferences depends only on the presence or absence of the incentive) or marginal (the effect depends on the extent of the incentive). We extend an earlier contribution (Bowles and Hwang, 2008) to include categorical crowding, thus providing a more general framework for the study of optimal incentives and as a result, an expanded range of situations for which the sophisticated planner will (surprisingly) make greater use of incentives when incentives crowd out social preferences than when motivational crowding is absent.

We thank Margaret Alexander, Lopamudra Banerjee, Ernst Fehr, Duncan Foley, John Geanakoplos, Bernd Irlenbusch, 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 two anonymous refrees for comments, and the Behavioral Sciences Program of the Santa Fe Institute and the U.S. National Science Foundation for support of this project.
Submitted: Nov 28 2011. **Revised:** September 15, 2012.

1 Introduction

The standard problem facing a social planner seeking to induce a target population to act more pro-socially (to contribute to a public good, for example) is to select an incentive, the effect of which will modify the material costs or benefits of the targeted activity so as to induce the desired behavior. But when incentives affect preferences, the sophisticated planner (aware of this complication) must also take account of the fact that incentives may adversely affect the targeted individuals' non-economic motivations that will also influence their actions, thus possibly reducing or even reversing the intended effect of the incentive. If the effect of the subsidy were separable in the incentives and the citizen's non economic preferences so that the use of an incentive had no effect on the target's pre-existing altruism or intrinsic pleasure in acting pro-socially, then the presence of these non-economic motivations would present no particular problem for the planner. But if the citizen's non-economic motivations are compromised by the implementation of the incentive, then the planner must take into account not only the direct effect of the incentives but also the possibly adverse indirect effects which occur when incentives crowd out social preferences. In this case incentives and social preferences are not separable, but instead are substitutes.

Both experiments and natural observations provide evidence of this non-separability of incentives and social preferences and suggest that the crowding out variant may be quite common (surveyed in Bowles (2008) and Bowles and Polania-Reyes(2012)). A growing empirical literature has explored the relationship between incentives and performance where the targets of the incentives are motivated by both conventional and social preferences (Bandiera et al., 2005; Bewley, 1999; Fehr and Goette, 2007; Fehr and Schmidt, 2007; Young and Burke, 2001). Moreover, a number of economic models have provided psychologically plausible mechanisms that may account for the crowding phenomenon (Benabou and Tirole (2006); Falk and Kosfeld (2006); Bar-Gill and Fershtman (2004)); for example, Bar-Gill and Fershtman (2005) provide a model in which a subsidy crowds out altruism by triggering an endogenous preference change. Finally, the neurological pathways whereby economic incentives may diminish pro-social behavior are beginning to be identified (Li et al., 2009).

In contrast to this literature, here we do not address the extent or causes of non-separability but instead like Frey (1999), Diamond (2006) and Funfgelt and Baumgartner (2012) investigate the implications for the optimal use of explicit economic incentives (which we will call simply incentives). We are particularly interested in whether incentives should be used less when they crowd out social preferences, as is commonly thought. (Crowding in is also observed in experiments, and we address this case as well as the more common crowding out case below).

A critical distinction that was absent from our earlier work (Bowles and Hwang, 2008) is that between a categorical form of crowding out, in which the mere presence of the incentive adversely affects social preferences, and a marginal form of crowding out in which the citizen's non-economic motivations to contribute vary inversely with the level of the incentive. We will see that categorical and marginal crowding out have different effects on optimal incentives and that in the presence of categorical crowding out (the case we did not address in our earlier work), the sophisticated planner will generally make more extensive use of incentives by comparison with the naive planner who thinks that preferences and incentives are separable. In this sense our earlier work is incomplete for if the sophisticated planner considers only

marginal crowding out she will miss a possibly large class of cases in which crowding out calls for more use of incentives rather than less. The current paper differs from our 2008 model additionally because we here consider an alternative planners' objective function in which the planner takes account of the effects of motivational crowding out on the citizen's behavior but otherwise abstracts from the incentive's negative effects on the citizen's pleasure in giving. We explain these alternative planners' objective functions in Section 3, and show in the concluding section that our main results hold for both objective functions.

2 Categorical and marginal crowding out: evidence

Because of the importance of the categorical-marginal distinction we begin with a recent experiment that allows an estimate of both categorical and marginal crowding. Irlenbusch and Ruchala (2008) implemented a public goods experiment in which subjects faced three conditions: no incentives to contribute and a bonus given to the highest contributing individual that was either high or low. Payoffs in the games were such that even with no incentive individuals would maximize their income by contributing 25 units. In the no-incentive case contributions averaged 37 units, or 48 percent above what would have occurred if the participants had been motivated only by the material rewards of the game, suggesting substantial effects of social preferences. Contributions in the low-bonus case were not significantly different from the no-bonus treatment. In the high-bonus case, however, significantly higher contributions occurred, but the amount contributed (53 units) barely (and insignificantly) exceeded that predicted for self-interested subjects (50 units). Thus while the high incentive "worked" (it increased contributions 43 percent over the no incentive case) it appears to have done this by substituting own-income-maximizing preferences for social preferences.

Bowles and Polania-Reyes(2012) estimated the marginal effect of the bonus in the Irlenbusch-Ruchala experiment using the observed behavior in the high and low bonus case along with the assumption that marginal crowding affects the slope of the citizens' best response function by a given amount (so that the function remains linear). The results are in Figure 1. Comparing the high and low bonus cases, they found that a unit increase in the bonus is associated with a 0.31 increase in contributions. This contrasts with the marginal effect of 0.42 that would have occurred under separability, that is, had own-income-maximizing subjects simply best responded to the incentive. Crowding out thus affected a 26 percent reduction in the marginal effect of the incentive. The estimated response to the incentive also gives us the level of categorical crowding out, namely the difference between the observed contributions in the absence of any incentive (37.04) and the predicted contributions had an arbitrarily small incentive been in effect (the vertical intercept of the "observed" line in Figure 1) or 34.55. The incentive thus categorically crowded out 21 percent of the effect of social preferences (measured by the excess in contribution levels above prediction for self-interested subjects, 12.04.)

Categorical crowding out is also evident in three experiments by Heyman and Ariely (2004). For example, reported willingness to help a stranger load a sofa into a van was much lower under a small money incentive than with no incentive at all, yet a moderate incentive increased the willingness to help (over the no incentive condition). Using these data as they did for the Irlenbusch and Ruchala study, Bowles and Polania-Reyes estimated that the mere

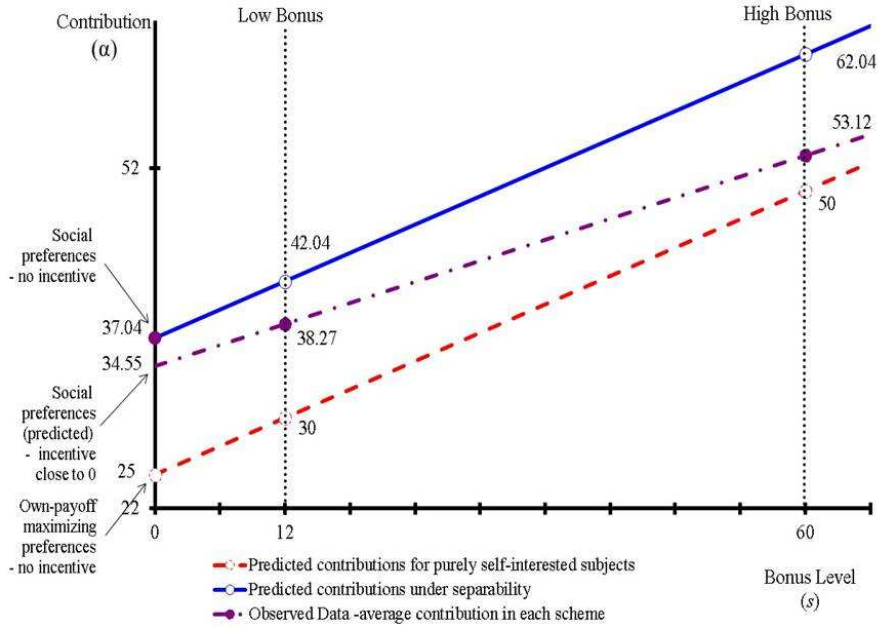

Figure 1: **Categorical and Marginal Crowding Out.** Source: Calculations by Bowles and Polania Reyes (2012) based on experimental results in Irlenbusch and Ruchala (2008)

presence of the incentive reduced the willingness to help by 27 percent (compared to the no incentive condition).

3 The effect of incentives on individual choice and Nash equilibrium contributions

To formalize these categorical and marginal incentive effects we need to model the influence of incentives on preferences. Incentives may change preferences, or may alter the motivational salience of a variety of pre-existing preferences. The distinction may be clarified in the Irlenbusch-Ruchala set-up: If a group of individuals accustomed to interacting under conditions similar to their high bonus were to occasionally find themselves in a no bonus situation, would they contribute the self regarding 25 units, because the exposure to the high bonus situation had given them self regarding preferences? Or would they contribute 37 units, because in the no bonus situation individuals would consider acting on self-regarding preferences to be inappropriate. Here we model the second case, namely, where incentives temporarily alter the salience of an individual's heterogeneous motives, because the exposure to the high bonus situation had given them self regarding preferences. In this case we say that preferences are state dependent, with the presence or extent of the incentive defining distinct states[1].

The key psychological insight captured by our model is that individuals have a multiplicity

[1]In the first case above incentives affect the process by which new preferences are learned, so that preferences are endogenous rather than state dependent. We address this case in Hwang and Bowles (2012).

of motives – self regarding, altruistic, spiteful and so on – the behavioral salience of which varies with the situation (Ross and Nisbett, 1991). Incentives affect preferences in this case because they alter the situation, providing cues as to whether the setting is more like, say, shopping, or like interacting with a close friend or family member. A psychologist would call these preferences situation-dependent, with incentives constituting the situation. Thus the preference function is an evaluation of states that the individual's action may bring about that is itself dependent on the current situation of the individual, and the latter varies according to the nature of incentives present. The other mechanism by which incentives may affect preferences – endogenous preferences– involves learning, resulting in a durable (not temporary) change in the preference function itself so that an individual who has learned new preference will subsequently behave differently in a given state.

To model the effects of state-dependent preferences on optimal incentives, consider a community of identical individuals indexed by $i = 1, ..., n$ who may contribute to a public project an amount $a_i \in [0,\ 1]$. The total contributions, $\sum_j a_j$, result in a benefit to each citizen of $\phi(\sum_j a_j)$, where $\phi(0) = 0$ and $\phi' > 0$. Each individual experiences the cost of contribution $g(a_i)$ where $g(0) = g'(0) = 0$, $g' > 0$ and $g'' > 0$. Incentives are provided in the form of a subsidy s that is proportional to the amount contributed. (By "incentive" we mean an intervention intended to influence an individual's behavior by altering the economic costs or benefits of some targeted activity).

To model the crowding problem we need to distinguish between an individual's social preferences in the absence of incentives and the preferences what will account for his behavior. The two will of course be identical in the absence of incentives but unless separability holds will be different otherwise. The former is a latent motivation which is here taken as exogenous and termed baseline social preferences (or just social preferences where no ambiguity will result.) The latter are realized preferences that vary with the presence and level of the incentives and that determine the citizen's behavior and we term these values.

We implement this distinction by assuming that depending on the incentives in force each citizen has "values" v that motivate pro-social behaviors (The relevant values include ethnical norms, a positive valuation on the well-being of others, intrinsic pleasures of cooperation *per se*, and other motives. But taking account of the multi-dimensional nature of the relevant values would not illuminate the question we address here). Thus citizen i's utility is

$$u(a_i) = \phi(\sum_j a_j) - g(a_i) + sa_i + v(s)a_i. \tag{1}$$

which makes it clear that the total effect of the subsidy on the citizen's action is a direct effect operating via the net costs of the targeted action (the penultimate term on the right hand side) plus a possible indirect effect operating via the influence of the subsidy on citizen's values (the final term on the right hand side).

To characterize the effects of categorical crowding and marginal crowding in terms of parameters, we adopt the following functional forms for the value function:

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

where $\lambda_0 \geq 0$ is the individual's baseline social preference motives to contribute to the public

good in the absence of an incentive, the indicator variable $\mathbf{1}_{\{s>0\}}$ takes the value of 1 when $s > 0$ and zero otherwise, and the "crowding parameters" λ_c and λ_m are the categorical and marginal effects of incentives representing the state dependent nature of preferences (preferences would be endogenous if variations in s affected these or other parameters of the value function, a case we do not address here).

From (1) and (2) we obtain citizen i's best response $a_i^{BR} := a_i^{BR}(a_{-i}, s)$ which is implicitly given by

$$g'(a_i^{BR}) = \phi'(a_i^{BR} + \sum_{j \neq i} a_j) + s + \lambda_0(1 + \mathbf{1}_{\{s>0\}}\lambda_c + s\lambda_m). \tag{3}$$

The last two terms in the right hand side of (3) show that the introduction of a subsidy increases contributions by raising the marginal benefits of contributing which we denote,

$$\beta := s + \lambda_0(1 + \mathbf{1}_{\{s>0\}}\lambda_c + s\lambda_m).$$

Considering the case in which there initially is no incentive, the effect of the introduction of an incentive on the net benefits of contributing (expressed in discrete terms so as to be able to account for the discontinuity in the value function at $s = 0$) is

$$\left.\frac{\Delta\beta}{\Delta s}\right|_{s=0} = 1 + \lambda_0(\frac{\lambda_c}{\Delta s} + \lambda_m) \tag{4}$$

and is composed (as expected) of a direct effect and the indirect effect which will be negative in the case of crowding out, and larger in absolute value the greater are the baseline social preferences of the individual (λ_0). When the subsidy level is positive, we find

$$\left.\frac{\Delta\beta}{\Delta s}\right|_{s>0} = 1 + \lambda_0\lambda_m. \tag{5}$$

We likewise see that

$$\frac{\Delta\beta}{\Delta\lambda_0} = 1 + \mathbf{1}_{\{s>0\}}\lambda_c + s\lambda_m \tag{6}$$

which in the case of crowding out is declining in s. Equations (4), (5), and (6) make it clear that when λ_c and λ_m are negative, incentives and baseline social preferences are substitutes in providing the motivation to contribute to the public good: the marginal effect of each varies inversely with the level of the other. Thus in the presence of crowding out, the adverse effects of incentives will be greater for individuals with a high level of baseline social preferences (for entirely self-interested individuals there is nothing to crowd out), which is what is observed in experiments(Kessler, 2008; Bohnet and Baytelman, 2007). If instead the second term on the right-hand side of (4) or (5) is positive incentives and social preferences are complements.

Using (4) and (5) we say that

categorical crowding out obtains	if $\frac{\Delta\beta}{\Delta s} < 1$	for $s = 0$ and for sufficiently small Δs
marginal crowding out obtains	if $\frac{d\beta}{ds} < 1$	for $s > 0$

.

If the signs are reversed, we have categorical and marginal crowding in. Note that when $\Delta\beta/\Delta s < 1$, the total effect of the incentive is less than the direct effect (and conversely for the case of crowding in). Crowding out will not occur if λ_c and λ_m or λ_0 are zero (values are not state dependent, or they are absent). Strong crowding out holds if $\Delta\beta/\Delta s < 0$ which can occur if categorical crowding out is large relative to the size and marginal effect of the subsidy, or if the marginal effect is negative. Note that crowding out does not require that the effect of the incentive be the opposite of that intended, only that it be less than would be the case were λ_c and λ_m or λ_0 zero.

Because citizens are identical and the subsidy non-discriminatory, there will be a symmetric Nash equilibrium in which everyone contributes the same amount to the public project. We discuss the detailed conditions for the existence of such equilibrium in the Appendix. We thus can drop the individual subscript and using (3) note that in order to be a mutual best response, the Nash contribution a^* must satisfy

$$\phi'(na^*) - g'(a^*) + s + \lambda_0(1 + \mathbf{1}_{\{s>0\}}\lambda_c + s\lambda_m) = 0. \tag{7}$$

To study the stability of the Nash equilibrium we consider a myopic best response dynamic whereby citizens maximize their utility, conditional on the contributions of others in the previous period (see the Appendix). We find that the condition for the asymptotic stability of the Nash equilibrium is given by

$$n\,|\phi''(na^*)| - g''(a^*) < 0, \tag{8}$$

a condition that ensures that the series of reciprocal effects of the citizens' actions converges (Given the convex cost function, (8) is satisfied if the benefit function is concave or "not too convex" given the number of group members).

The public project's net benefit function, defined as $\phi(na) - g(a)$, plays an important role in what follows. We make the following assumptions on the net benefit function.

Assumptions

A1 In the absence of a subsidy, citizens under-contribute to the project: the net benefit function is increasing in the level of contribution (i.e., $n\phi'(n\underline{a}) - g'(\underline{a}) > 0$, where $\underline{a}$ denotes the contribution level without subsidy.)

A2 The Nash equilibrium a^* is asymptotically stable (i.e., $n\,|\phi''(na^*)| - g''(a^*) < 0$).

The sophisticated planner affects citizens' contributions by selecting s to implement a Nash equilibrium given by equation (7), implying the "implementation technologies" for alternative values of the crowding parameters illustrated in Figure 2: $a^*(s, \lambda_m, \lambda_c)$.

When the level of subsidy s is positive, using equation (7) we find expressions of the derivatives of a^* with respect to s, λ_c, λ_m as follows:

$$\frac{\partial a^*}{\partial s} = \frac{1 + \lambda_0\lambda_m}{g'' - n\phi''}\,, \qquad \frac{\partial a^*}{\partial \lambda_c} = \frac{\lambda_0}{g'' - n\phi''}, \qquad \frac{\partial a^*}{\partial \lambda_m} = \frac{s\lambda_0}{g'' - n\phi''}. \tag{9}$$

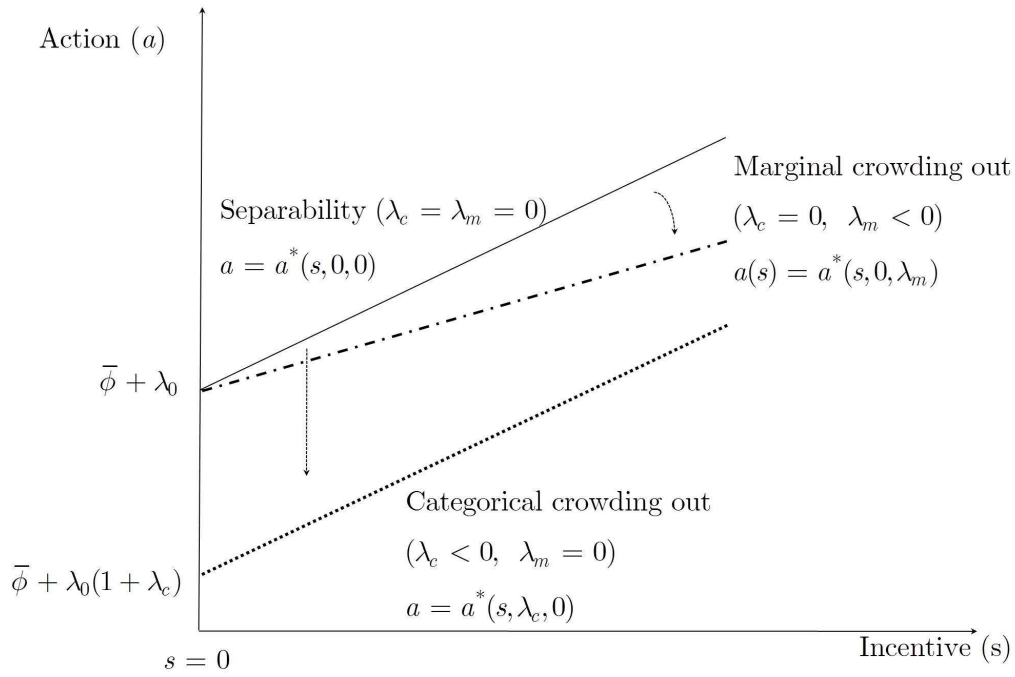

Figure 2: **The sophisticated planner's technology: citizen's Nash equilibrium action response function** $a^*(s, \lambda_c, \lambda_m)$**.** In the figure, we take $g(a) = \frac{1}{2}a^2$ and $\phi(\sum a_i) = \bar{\phi} \sum a_i$.

A2 ensures that the signs of $\partial a^*/\partial s$, $\partial a^*/\partial \lambda_c$, and $\partial a^*/\partial \lambda_m$ are all positive. Note that the effect of variations in the subsidy on the Nash contribution level depends not only on the sum of the direct incentive effect and its crowding effect $(1 + \lambda_0\lambda_m)$ but also (inversely) on the rate at which the individual's marginal net benefits diminish with increases in the contribution level $(g''(a) - n\phi''(na))$. Thus strongly increasing marginal costs of contribution or diminishing marginal returns to the project dampen the effect of the subsidy. The same is true of the effects of variations in the two crowding parameters, as the two other equations in (9) make clear.

Note from the definition of crowding in and out, equation (3) (the individual best response) and Figure 2 that crowding out may be present even in the absence of a negative effect of the subsidy on the individual's action or on the Nash equilibrium $(\partial a^*/\partial s)$. In the case of marginal crowding out, the effect of the incentive on the individual's action will be positive unless the marginal crowding out effect on values more than offsets the effect of incentives on the citizen's payoffs. In the case of categorical crowding out the effect on individual and Nash equilibrium contributions will be positive for a sufficiently large subsidy.

4 The planner's problem when preferences are state dependent

We now turn to the problem of a social planner who chooses a subsidy level to maximize the net benefits of the public project, taking account not only of the direct effect of the

incentive on the individual's private marginal benefits of contributing but also the effect of the subsidy on the citizen's preferences and thereby, indirectly on the Nash equilibrium contribution levels. We suppose that there is a cost per citizen of administering the subsidy $c(s)$ where $c'(0) = 0 = c(0)$, $c'(s) > 0$ and $c''(s) > 0$ for $s > 0$. Since the sophisticated planner is aware of non-separability, she correctly expects the citizens' Nash equilibrium in response to the subsidy to be $a^* = a^*(s, \lambda_c, \lambda_m)$ and selects s^*. Her naive counterpart is a planner who ignores non-separability assumes that no crowding-out obtains ($\lambda_c = \lambda_m = 0$); so to him the expected Nash equilibrium is simply $a^N = a^*(s, 0, 0)$, that is, the top line in Figure 2. Except where crowding is absent, the two planners will select different levels of subsidy. We say that the subsidy is overused by the naive planner if the subsidy he selects, s^N exceeds that selected by the sophisticated planner, s^*.

Because citizens are identical, the sophisticated social planner's optimizing problem can be reduced to maximizing the net benefits of the public project for a single citizen, and written as

$$\begin{aligned} &\max_{a\in[0,1],\ s\in[0,1]} \omega(a, s) := \phi(na) - g(a) - c(s) \\ &\text{subject to } \ a = a^*(s, \lambda_c, \lambda_m). \end{aligned} \tag{10}$$

To exclude uninteresting cases we assume that $c(s)$ is sufficiently convex so that there is a unique interior solution of (10) (See the Appendix). Notice we assume that the social planner treats the citizens' values as a component of individual motivation, but does not include them in her objective function (10). Thus, from a normative point of view we exclude from the planner's welfare function the citizen's subjective pleasure of contributing, and restrict the effect of the project on social welfare to its conventionally defined benefits and costs. See Bowles and Hwang (2008) for the alternative case in which the citizens' values are included not only in the citizens' best response functions but also in the planner' objective function. In this latter case, if crowding out obtains, incentives reduce the citizen's pre-existing pleasure of contributing, and the sophisticated planner takes account of this.

To investigate underuse and overuse of incentives by the naive planner, we define the marginal rate of substitution (σ) and the marginal rate of transformation (τ) :

$$\tau(a, \lambda_m) := \frac{1 + \lambda_0\lambda_m}{g''(a) - n\phi''(na)}, \quad \sigma(a, s) := \frac{c'(s)}{n\phi'(na) - g'(a)}. \tag{11}$$

The marginal rate of transformation, evaluated at the citizen's Nash response $a^*(s)$, is just the effectiveness of the subsidy in altering Nash contributions (from (9), the slope of the implementation functions). The marginal rate of substitution is the ratio of the planner's marginal costs of varying the subsidy to the marginal benefits to the planner of variations in the contribution level (the slope of planner's indifference loci). Then the first order condition for an interior equilibrium of the sophisticated planner's problem equates the marginal rate of transformation of the subsidy into the provision of public good with the marginal rate of substitution between the public good and the subsidy:

$$\tau(a(s, \lambda_c, \lambda_m), \lambda_m) - \sigma(a^*(s, \lambda_c, \lambda_m), s) = 0. \tag{12}$$

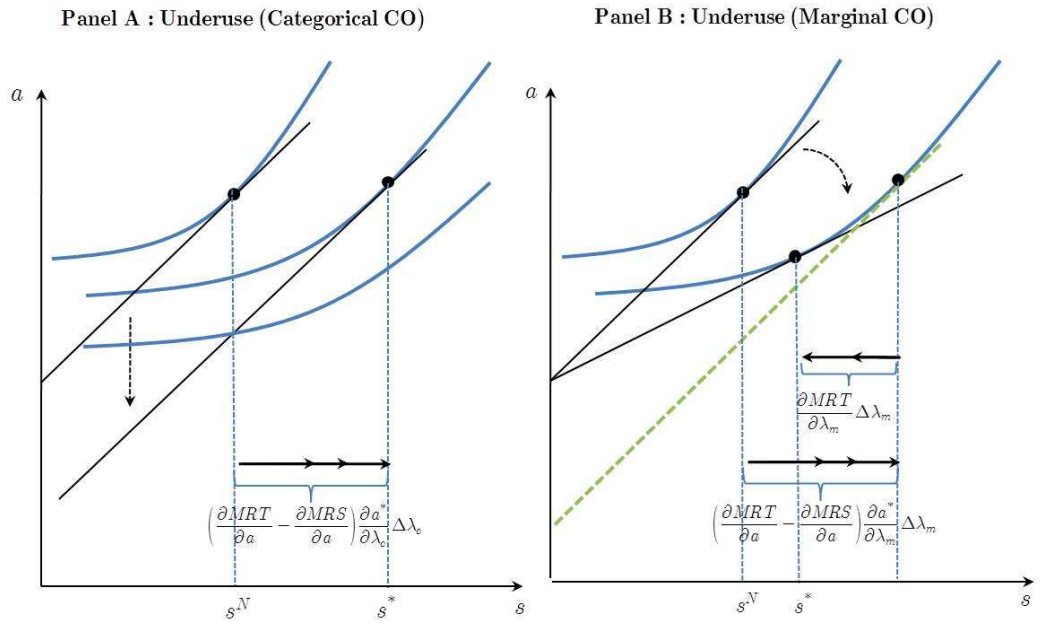

Figure 3: **The Naive Planner's Underuse of Incentives when Crowding out Occurs:Illustration of Proposition 1.** To illustrate the intuition behind categorical crowding out leading to underuse of the subsidy by the naive planner, the case shown assumes $\phi'' = 0 = g'''$so that the reduction in a occasioned by categorical crowding out does not alter the marginal rate of transformation. In the case of marginal crowding out, either underuse or overuse may occur. Panel B shows the joint effect of the reduction in the marginal rate of substitution between a and s (the flatter indifference locus shifting s^* to the right) and the reduced marginal rate of transformation between a and s. (the flatter implementation function) partially offsetting the first effect.

5 The effect of crowding out on optimal incentives

Crowding out will affect both the marginal rate of substitution and the marginal rate of transformation in (12). Consider first categorical crowding out and its effect on the optimal subsidy, taking the naive planner's subsidy s^N as the status quo. This shifts downward the planners' implementation technology, reducing the provision of the public good and thereby increasing or decreasing the marginal benefit of public good provision (if the benefit function is respectively concave or convex). The result is that at the given subsidy level (s^N) and resulting citizen's contribution level the planner's indifference locus will be flatter if the benefit function is concave (and steeper if it is convex) as shown in Panel A of Figure 3. We will see below that categorical crowding may (but need not) also affect the marginal rate of transformation, but in Figure 3 we show a case where it does not. In the case illustrated the benefit function is concave so the reduced public goods provision raises the marginal benefit of the public good, flattening the iso-benefit locus. This flattening of the planner's indifference loci, shown in Figure 3, Panel A displaces the optimal subsidy to the right, so that the presence of crowding out increases the optimal subsidy (from s^N to s^*).

Marginal crowding out has two effects. The first is the reduction in the marginal effectiveness of the subsidy, which reduces the marginal rate of transformation ("flattens" the

planner's implementation function, as shown in Panel B, Figure 3). Second, as in the case of categorical crowding out, the reduced provision of the public good alters the marginal rate of substitution, flattening the planner's indifference locus if the benefit function is concave. Which of these effects dominates determines the effect of crowding out on the optimal subsidy. In Panel B of Figure 5 we illustrate the case in which marginal crowding out increases the optimal subsidy because the heightened marginal benefit of the public good more than offsets the reduced effectiveness of the subsidy (If the benefit function is convex, marginal crowding out unambiguously induces a reduction in the sophisticated planner's s^* because both effects work in the same direction: the incentive is less effective and the marginal benefits of the project are reduced by the lesser level of provision.).

We can now formalize these intuitions. First we find the effect of categorical crowding on the optimal subsidy ($ds^*/d\lambda_c$). To do this, we totally differentiate (12) with respect to s and λ_c and find that

$$\frac{ds^*}{d\lambda_c} = -\frac{1}{(\frac{\partial\tau}{\partial a} - \frac{\partial\sigma}{\partial a})\frac{\partial a^*}{\partial s} - \frac{\partial\sigma}{\partial s}} \left[\frac{\partial\tau}{\partial a} - \frac{\partial\sigma}{\partial a}\right] \frac{\partial a^*}{\partial\lambda_c}. \tag{13}$$

The second order condition for the social planner's maximization problem requires that $c(s)$ is sufficiently convex and thus the denominator of (13) (which is just the derivative of the left hand side of (12) with respect to s) is negative. Note that when categorical crowding out occurs (i.e., λ_c decreases from 0 to a negative value), the contribution level will be reduced ($\partial a^*/\partial\lambda_c > 0$), so the sign of $ds^*/d\lambda_c$ depends on $\partial\tau/\partial a - \partial\sigma/\partial a$.

Concerning marginal crowding, similarly we find

$$\frac{ds^*}{d\lambda_m} = -\frac{1}{(\frac{\partial\tau}{\partial a} - \frac{\partial\sigma}{\partial a})\frac{\partial a^*}{\partial s} - \frac{\partial\sigma}{\partial s}} \left[\frac{\partial\tau}{\partial\lambda_m} + (\frac{\partial\tau}{\partial a} - \frac{\partial\sigma}{\partial a})\frac{\partial a^*}{\partial\lambda_m}\right] \tag{14}$$

which using using (9) and (11) can be rewritten:

$$\frac{ds^*}{d\lambda_m} = -\frac{1}{(\frac{\partial\tau}{\partial a} - \frac{\partial\sigma}{\partial a})\frac{\partial a^*}{\partial s} - \frac{\partial\sigma}{\partial s}} \left[1 + (\frac{\partial\tau}{\partial a} - \frac{\partial\sigma}{\partial a})s^*\right] \frac{\lambda_0}{g'' - n\phi''} \tag{15}$$

Since the denominator of this expression is negative and the final term positive (by the condition for the stability of the Nash equilibrium), $ds^*/d\lambda_m$ will have the same sign as the terms in the square bracket in (15). Using (13) and (15) we find:

Proposition 1 ***(Underuse of Incentives by the Naive Planner when Incentives Crowd out Preferences)*** *Suppose that A1 and A1 hold and $s^* > 0$. Then*

$$\begin{aligned}
\frac{ds^*}{d\lambda_c} &< 0 \qquad \textit{if and only if} \quad \frac{\partial\sigma}{\partial a}(a^N, \lambda_m) - \frac{\partial\tau}{\partial a}(a^N, s^N) > 0 \\
\frac{ds^*}{d\lambda_m} &< 0 \qquad \textit{if and only if} \quad \frac{\partial\sigma}{\partial a}(a^N, \lambda_m) - \frac{\partial\tau}{\partial a}(a^N, s^N) > \frac{1}{s^N}
\end{aligned}$$

Notice that Proposition 1 provides the characterization of marginal and categorical crowding "locally"; it asserts that whenever the conditions on τ and σ hold at the naive planner's

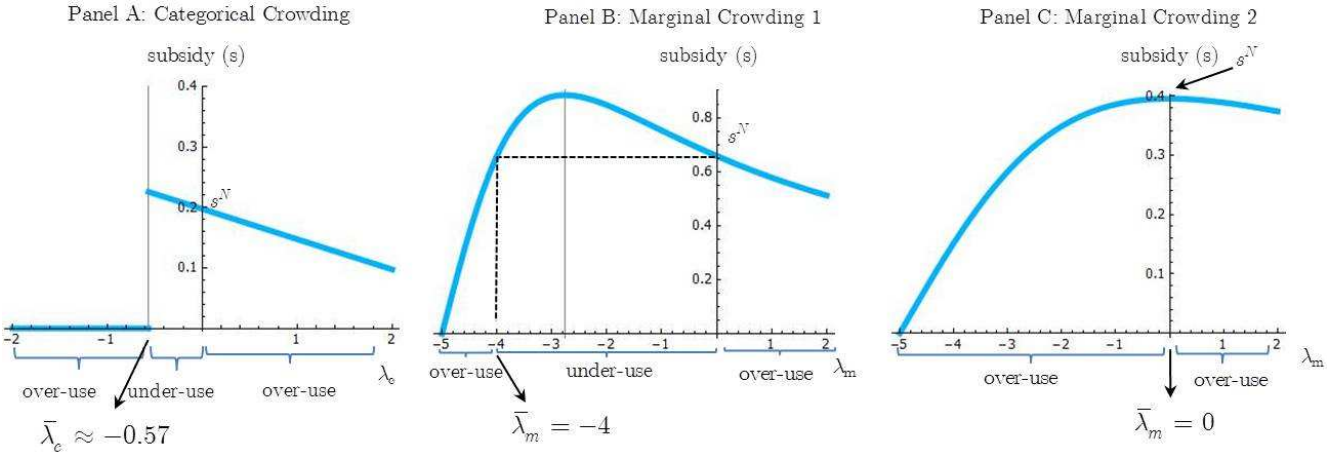

Figure 4: **Sophisticated planner's optimal subsidy and underuse of incentives by the planner under crowding out.** Negative (positive) values of λ_c and λ_m imply crowding out (in). In the case of the categorical crowding out (Panel A), the optimal subsidy increases as λ_c decreases (meaning stronger categorical crowding out) until at $\lambda_c \approx -0.57$, the social planner will set $s^* = 0$. Panels B and C show the choices of the sophisticated planners in the case of marginal crowding, depending on different values of the paramenters. In Panel B in the case of crowding out (negative values of λ_m) for any level of marginal crowding out greater than -4, the sophisticated planner implements are larger subsidy than the naive planner. In Panels B and C, when $\lambda_m = -5$ the sophisticated planner abandons use of the subsidy entirely because then $\lambda_0 \times \lambda_m = -1$ and equation (**??**) shows that the subsidy has no effect on the citizens' actions while larger negative values of λ_m imply what we have defined as strong crowding out (the effect is opposite of that intended). For three panels: $\phi(\sum a_i) = 0.11 \sum a_i$, $\lambda_0 = 0.2$. Panel A: $\gamma = 3, \kappa = 1$, $\lambda_m = -0.2$, Panel B: $\gamma = 1$, $\kappa = 0.2, \lambda_c = 0$, Panel C: $\gamma = 1, \kappa = 1$, $\lambda_c = 0$. See the Appendix for the definitions of parameters and analysis.

choice, a small decrease in λ_c (or λ_m) would raise the subsidy level selected by the sophisticated planner. Therefore if the conditions in Proposition 1 hold at $\lambda_c = \lambda_m = 0$, we may conclude that the underuse of incentives would occur if a small degree of crowding out obtains (See also Figure 4).

The proposition shows that if categorical crowding out reduces the marginal rate of substitution (at s^N) by more than it reduces the marginal rate of transformation it will induce the sophisticated planner to make greater use of incentives than her naive counterpart. In the case of marginal crowding out, the former effect must exceed the latter by $1/s^N$. This is because if s^N is very small the change in the slope of the implementation function will dominate as a result of the fact that the reduction is contributions will be small (because with marginal crowding the implementation function does not shift down, it rotates clockwise), and hence the increase in the marginal benefit of the public good will be insufficient to offset the decreased effectiveness of the subsidy.

Using (13) and (15) we also find

$$\frac{ds^*}{d\lambda_m} = \underbrace{-\frac{1}{(\frac{\partial\tau}{\partial a} - \frac{\partial\sigma}{\partial a})\frac{\partial a^*}{\partial s} - \frac{\partial\sigma}{\partial s}}\frac{\partial a^*}{\partial \lambda_c}}_{>0} + s^*\frac{ds^*}{d\lambda_c}.$$

Proposition 2 ***(Categorical Crowding out vs Marginal Crowding out)*** *Suppose that A1 and A2 hold and $s^* > 0$. The condition under which categorical crowding out leads to the underuse by the naive planner is less stringent than the condition under which marginal crowding out leads to the underuse:i.e.,*

$$\frac{ds^*}{d\lambda_m} < 0 \text{ implies } \frac{ds^*}{d\lambda_c} < 0.$$

Proposition 2 shows that the conditions for the parameters under which categorical crowding out leads to the underuse of incentives by the naive planner are more general than those for marginal crowding out and this suggests that underuse phenomena occurs more widely under categorical crowding out than marginal crowding out. This is because compared to the case of the categorical crowding, the marginal rate of transformation is directly affected by λ_m (the effect captured by $\partial\tau/\partial\lambda_m$). Therefore for $ds^*/d\lambda_m$ to be negative (leading to underuse of incentives by the naive planner), the flattening effect of crowding out on the indifference locus must be greater than the case of categorical crowding if it is to offset the direct effect of flattening of the best responses (See Figure 3 Panel B).

In Figure 4, to illustrate the economic intuitions underlying our results, we show categorical and marginal crowding out, adopting simple functional forms. In Panels A and B moderate crowding out induces the sophisticated planner to adopt a larger subsidy than her naive counterpart, indicating underuse of the incentive by the naive planner. But there is a critical level of crowding out which if exceeded induces her either to abandon the incentive altogether (Panel A) or to adopt a lesser incentive than her naive counterpart. Notice that (as in Panel C of Figure 4) the sophisticated planner may adopt a lower subsidy than the naive planner (the naive planner overuses the incentive) irrespective of whether non-separability takes the form of crowding in or crowding out. This, and the non monotonicity of s^* in λ_m in Panel B, occurs because the sign of the terms in brackets in (15) changes due to variations in contribution levels resulting from the variation in the crowding parameter. In panel C we have chosen parameters such that the expression changes sign at $\lambda_m = 0$.

6 Extensions and modifications: Taxes, additional instruments, and the utilitarian planner

The use of incentives to promote such pro-social behavior as contributing to a public good may compromise non-economic motivations that would have induced an individual to act in pro-social ways in the absence of the incentive. We have distinguished between a marginal and a categorical form of this motivational crowding out process and characterized the optimal use of incentives in the presence of each. We have shown that crowding out may induce

the sophisticated planner to make greater use of incentives that would her naive counterpart who is unaware of the crowding phenomenon, and that the conditions under which this underuse of the incentive by the naive planner are more general when crowding is of the categorical type than when only marginal crowding out occurs. The reason is that unlike marginal crowding out, categorical crowding out does not reduce the marginal effectiveness of the incentive; it simply reduces the level of contribution to the public good, thereby raising the marginal payoff to policies promoting contributions, and possibly inducing the sophisticated planner to make greater use of the incentive.

A natural question is to ask whether the same analysis applies to taxes. The simple answer is yes: the model is entirely unaffected by considering a tax on the citizen's shortfall from the socially optimal contribution level $a = 1$ rather than a subsidy on her contribution level. As an empirical matter, taxes might have a greater crowding out effect than subsidies if the mechanism by which crowding occurs is that the incentive conveys adverse information about the individual designing the incentive - penalties being more likely to convey this kind of 'bad news' than rewards (Evidence for this "bad news" interpretation of the crowding out phenomenon is surveyed in Bowles and Polania-Reyes, 2012.)

How robust is the result that the sophisticated planner may make greater use of incentives than the naive planner? The underuse of incentives by the naive planner arises because the net benefit function is concave (so that shortfalls from the optimal provision of the public good are increasingly costly in welfare terms), and crowding out leads to a larger shortfall than would occur in its absence. If the only instrument available to the planner is the subsidy, then a sufficiently concave net benefits function will induce her to adopt a larger subsidy. But suppose the planner were to have instruments for the promotion of contributions to the public good other than incentives. She might, for example, promote publicity making the citizen's duty to contribute more salient, as in Frey (1999). In this case marginal crowding out would induce the sophisticated planner not only to devote more resources to enhancing contributions to the public good but also to substitute the alternative instrument for the subsidy. Thus the conditions for underuse of the incentive by the sophisticated planner (Proposition 1) are more stringent when there exists an alternative instrument that is not subject to crowding out.

To consider a final extension, recall that our planner maximizes the conventionally measured net benefits of the public project, taking account of the behavioral responses of the citizens induced by their social preferences but disregarding the possibly negative effect of incentives on the citizen's ethical or intrinsic pleasure of contributing. But ought the planner instead to consider the citizen's ethical values hedonistically, treating them as simply another form of 'tastes,' the satisfaction of which is pleasurable to the citizen and which the planner should consider relevant in her evaluation of the outcomes of her policies. Like others (for example, Diamond (2006) and Bergstrom (2006)) we question whether the preferences that explain individual behavior should necessarily constitute planner's objective 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 at the same time providing a basis for the evaluation of the outcomes of the policies under study. It is far from obvious that the same concept can perform both tasks. Diamond (2006) after having presented the citizen's response to incentives based on the

individual's utility function (including warm glow altruism) remarks (correctly, we think): "That behavior is describable in this way does not necessarily imply that social welfare should be defined in the same way." (p.909) The two uses of utility often coincide: we think that the citizen's enjoyment in eating fruit is part of what the planner should maximize. But what about the obese citizen's pleasure in consuming sweets? The utility functions used to predict behaviors may require taking account of addictions, hyperbolic discounting, weakness of will and other empirically observed aspects of motivation. Does that also require that we value the satisfaction of these often self destructive preferences when considering what the planner should optimize? We do not think so.

In cases where these motives lead to self-destructive behavior, the case for not treating their satisfaction as part of the normative standard for policy evaluation is convincing. If some treatment for drug addiction, for example, were to reduce the addict's exhilaration when shooting up, we would not count this reduced exhilaration as a cost of the program. A thoroughgoing utilitarian planner, however, might count these subjective effects as part of her objective function when evaluating the drug policy.

Addressing this philosophically complex question adequately in this brief note is not possible, and without resolving the issue we think that our 'materialist' planner's objective function is a plausible formulation. But one wonders if similar results would obtain under what we will term a 'utilitarian' planner's objective function. Fortunately the use of either of the two planners' objective function does not alter our qualitative results.

One's intuition is that the effects of non-separability would be exaggerated in the case of the utilitarian planner. In the case of marginal crowding out, for example, the sophisticated planner would have a second reason for minimizing the use of incentives: not only would the incentive be less effective in inducing the citizen to contribute (the reason studied above) but it would also reduce the citizen's utility derived from the value of giving. For the case of marginal crowding out studied in our 2008 paper, this intuition is confirmed. But nonetheless we find that for a sufficiently concave net benefits function the sophisticated utilitarian planner will make greater use of the subsidy than her naive counterpart.

To see that this result holds also for the utilitarian planner in the case of categorical crowding out, recall that in the case of the materialist planner studied in this paper the conditions under which categorical crowding out will lead to underuse of the incentive by the naive planner are less stringent than the case of marginal crowding (Proposition 2). The reason is that while both forms of crowding reduce public goods provision and raise the marginal net benefits of provision, unlike marginal crowding out, under categorical crowding out there is nothing to offset the reason for the sophisticated planner's greater use of the incentive by reducing the marginal effectiveness of the subsidy. In the appendix we show that the same holds for the utilitarian planner, with the result that the fact (from our 2008 paper) that the utilitarian planner may underuse the incentive in the presence of marginal crowding out holds *a fortiori* in the case of categorical crowding out.

7 Conclusion

The sophisticated planner now knows that incentives and social preferences need not be separable but may be either complements or (more likely) substitutes, that both categorical

and marginal crowding effects may occur, that she may be able to estimate their magnitude on the basis of experiments, and that taking account of crowding out effects may induce her to adopt either greater or lesser incentives than would have been the case had she remained unaware of the non-separability problem. The curious cases in which her naive counterpart would underuse incentives in the presence of crowding out will occur (unsurprisingly, she now realizes) when there are strong diminishing net returns to the public project. She wonders why she did not learn any of this in school.

There were two things missing from the standard model, she recalls. First, because preferences were assumed to depend only on one's own material payoffs, there were no social preferences to crowd in or out. And second (a more subtle point): if there were social preferences relevant to the problem under analysis they were assumed (implicitly) to be just additive with any payoff-based incentives that the planner might provide. This separability assumption appeared natural because in the standard model preferences were evaluations of states, consumption bundles, for example, defined simply as vectors of commodities that one might consume. The reason why some states were feasible and others not was a matter of the budget constraint, and had no effect on preferences (social or otherwise). Other than its effect on the budget constraint, a vector of goods acquired by purchase was, in this framework, no different from one acquired directly by one's own labor, from a charity, or from state as a citizen's right.

There is something wrong with this picture, the planner now realizes: people care about the processes by which states come to be in their feasible set. The most plausible psychological foundation for what we will call processes-dependent preferences is that when people take an action it may be to acquire something (as in many economic actions) but often it is also because the person wants to be or to become a certain kind of person in one's own eyes or in the eyes of others (Cooley, 1902; Leung and Martin, 2003; Akerlof and Kranton, 2010). For example, the desire not to be (and be seen as) a chump may explain why experimental subjects respond very differently to being offered a disadvantageous share of a pie in an Ultimatum Game depending on whether the share was determined by another subject or by a computer (Blout, 1995). A disadvantageous offer from a computer is just bad luck and tends to be accepted as preferable to no offer at all; but the same offer from another subject signals unfairness of the proposer and tends to be rejected. In economic situations preferences appear to be process-dependent for two reasons; the process by which a state comes to be on one's feasible set often reveals important information about the intentions of others, and also provides cues concerning socially appropriate behaviors. Responses to incentives indicating non-separability are simply a well documented case of this more general class.

Were we to index states by the kinds of incentives associated with their being feasible, the planner muses, we might have a model of economic behavior in which non-separability of material incentives and social preferences could occur naturally. This would entail treating states not simply as vectors of things to have, but as activities, that is, something one does. She then remembers that Kelvin Lancaster had long ago proposed a "new approach to consumer theory" along just these lines. The paper is still in her filing cabinet. She reads: "The good, *per se*, does not give utility to the consumer..." then, "Consumption is an activity in which goods are inputs and in which the output is a collection of characteristics" (Lancaster (1966):133-4). Like Lancaster, the planner might reconstruct the fundamentals of

economic behavior by representing consumption as just another form of production, in which the object of production would include the actor's self esteem, standing in the community, and other non-material objectives. But doing this would take the planner away from her job, and take the authors of this paper beyond the limits of this brief note.

References

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

Bandiera, O., I. Barankay, and I. Rasul (2005). Social preferences and the response to incentives: Evidence from personnel data. *The Quarterly Journal of Economics 120:3*, 917–62.

Bar-Gill, O. and C. Fershtman (2004). Law and preferences. *Journal of Law, Economics and Organization 20*(2), 331–53.

Bar-Gill, O. and C. Fershtman (2005). Public policy with endogenous preferences. *Journal of Public Economic Theory 7 (5)*, 841–857.

Benabou, R. and J. Tirole (2006). Incentives and prosocial behavior. *American Economic Review 96*(5), 1652–78.

Bergstrom, T. (2006). Benefit-cost in a benevolent society. *American Economic Review 96*(1), 339–51.

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

Blout, S. (1995). When social outcomes aren't fair: The effect of caual attributions on preferences. *Organizational Behavior and Human Decision Processes 63*(2), 131–44.

Bohnet, I. and Y. Baytelman (2007). Institution and trust - implications for preferences, beliefs, and behavior. *Rationality and Society 19*, 99–135.

Bowles, S. (2008). Policies designed for self-interested citizens may undermine "the moral sentiments:" evidence from experiments. *Science 320*(5883), 1605–9.

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

Bowles, S. and S. Polania Reyes (2012). Economic incentives and pro-social behavior. *Journal of Economic Literature (forthcoming)*.

Cooley, C. H. (1902). *Human Nature and the Social Order*. New York: Charles Scribner's Sons.

Diamond, P. (2006). Optimal tax treatment of private contributions for public goods with and without warm glow preferences. *Journal of Public Economics 90*, 897–919.

Falk, A. and M. Kosfeld (2006). The hidden costs of control. *American Economic Review 96*(5), 1611–30.

Fehr, E. and L. Goette (2007). Do workers work more if wages are high? evidence from a randomized field experiment. *American Economic Review 97:1*, 298–317.

Fehr, E. and K. M. Schmidt (2007). Adding a stick to the carrot? the interaction of bonuses and fines. *American Economic Review 97:2*, 177–81.

Frey, B. S. (1999). Morality and rationality in enviromental policy. *Journal of Consumer Policy 22*, 395–417.

Funfgelt, J. and S. Baumgartner (2012). Regulation of morally responsible agents with motivation crowding. Leuphana University of Lneburg.

Heyman, J. and D. Ariely (2004). Effort for payment: A tale of two markets. *Psychological Science 15*, 787–93.

Hwang, S.-H. and S. Bowles (2012). The sophisticated planner's dilemma: optimal incentives with endogenous preferences. Santa Fe Institute Working Paper.

Irlenbusch, B. and G. Ruchala (2008). Relative rewards within team-based compensation. *Labour Economics 15*, 141–67.

Kessler, E. (2008). *Behavioral Economics of Performance Incentives*. Ph. D. thesis, School of Economics, University of Nottingham.

Lancaster, K. (1966). A new approach to consumer theory. *Journal of Political Economy 74*(2), 132–57.

Leung, K. T. and J. L. Martin (2003). The looking-glass self: An empirical test and elaboration. *Social Forces 81*(3), 593–622.

Li, J., E. Xiao, D. Houser, and P. R. Montague (2009). Neural responses to saction threats in two-party economic exchanges. *Proceedings of the National Academy of Science 106*(39), 16835–16840.

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

Young, P. and M. A. Burke (2001). Competition and custom in economic contracts: A case study of illinois agriculture. *American Economic Review 91*, 559–573.