A note on optimal incentives with state-dependent preferences

Sung-Ha Hwang\textsuperscript{a}, Samuel Bowles\textsuperscript{b,*}

\textsuperscript{a}Department of Mathematics and Statistics, University of Massachusetts at Amherst, MA, 01003-9305, U.S.A.

\textsuperscript{b}Santa Fe Institute, U.S.A. and Dipartimento di Economia Politica, University of Siena, Italy

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 payoff maximizing motives may crowd out non-economic motives such as altruism, reciprocity, intrinsic motivation and other social preferences, leading to disappointing and sometimes even counter-productive incentive effects. Evidence from behavioral experiments indicates 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 to this journal (Bowles and Hwang, 2008) providing a more general framework for the study of optimal incentives when crowding out results from framing and information effects including (with evidence for ) categorical crowding, and as a result, an expanded range of situations for which the sophisticated planner will make greater use of incentives when incentives crowd out social preferences than when motivational crowding is absent.

Keywords:
Social preferences, public goods, motivational crowding out, explicit incentives, framing, endogenous preferences

JEL Classification Numbers: D64 (Altruism); H41 (Public goods); D78 (Policy making and implementation); D03 (Behavioral Economics)

\*Corresponding author, Santa Fe Institute, 1399 Hyde Park Rd., Santa Fe, NM, 87501, U.S.A. Phone: 1-505-984-8800 Fax:1-505-982-0565

Email addresses: h wang@math.umass.edu (Sung-Ha Hwang), samuel.bowles@gmail.com (Samuel Bowles)
1. Introduction

When incentives affect preferences, the problem facing the sophisticated 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 not only be (as intended) to modify the material costs or benefits of the targeted activity but also possibly to adversely influence the targeted individual’s non-economic preferences that will also affect their actions. If the behavioral effect of the subsidy were separable in the incentives and the citizen’s non economic preferences, the presence of the latter 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 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 of it may be quite common (surveyed in Bowles (2008) and Bowles and Polania Reyes(2011)). 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)) and the neurological pathways whereby economic incentives may diminish pro-social behavior are beginning to be identified (Li et al., 2009)). Here we do not address the extent or causes of non-separability but rather 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.

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 (Categorical and marginal crowding in is also observed in experiments, and we address these cases as well as the more common crowding out below). We will see that categorical and marginal crowding out have very 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.
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 were such that even with no incentive individuals would maximize their payoffs by contributing 25 units. In the no-incentive case contributions averaged 37 units, or 48 percent above what would have occurred if the participants had been motivated only by the material rewards of the game, 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 payoff-maximizing preferences for social preferences.

Bowles and Polania Reyes (2011) estimated the marginal effect of the bonus 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 payoff 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 Ariely et al. (2009). 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 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 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.\(^1\)

\(^1\)In the first case above incentives affect the process by which new preferences are learned, so that preferences are endogenous. We address this case in Hwang and Bowles (2011).
The key psychological insight captured by our model is that individuals have a multiplicity 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. We use the term state-dependent for this case, where the presence and level of an incentive is an aspect of the state. 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 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 \( \phi(\sum_j a_j) \), while each individual experiences the cost of contribution \( g(a_i) \) where \( \phi(0) = 0, \phi' > 0, g(0) = g'(0) = 0, g' > 0 \) and \( g'' > 0 \). Incentives are provided in the form of a subsidy \( s \) 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 her behavior. The two will 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 endogenous realized preferences 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 \( \text{per se} \), 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

\[
u(a_i) = \phi(\sum_j a_j) - g(a_i) + sa_i + v(s)a_i.\] (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 plus a possible indirect effect operating via the influence of the subsidy on citizen’s values.

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

\[ v(s) = \lambda_0(1 + 1_{s>0})\lambda_c + s\lambda_m) \]  

(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 \( 1_{s>0} \) takes the value of 1 when \( s > 0 \) and zero otherwise, and the “crowding parameters” \( \lambda_c \) and \( \lambda_m \) are the categorical and marginal effects of incentives representing the state dependent nature of preferences (preferences would be endogenous if variations in \( s \) affected any of the parameters of the value function, a case we do not address here).

With this setting, citizen \( i \)'s maximization problem is given by

\[
\max_{a_i \in [0,1]} u(a_i) := \phi\left( \sum_j a_j \right) - g(a_i) + sa_i + \lambda_0(1 + 1_{s>0})\lambda_c + s\lambda_m)a_i.
\]  

(3)

From (3) we obtain citizen \( i \)'s best response \( a_i^{BR} := a_i^{BR}(a_{-i}, s) \) which is implicitly given by

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

(4)

The last two terms in the right hand side of (4) show that the introduction of a subsidy increases contributions by raising the marginal benefits of contributing which we denote,

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

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

\[
\frac{\Delta \beta}{\Delta s}\bigg|_{s=0} = 1 + \lambda_0(\frac{\lambda_c}{\Delta s} + \lambda_m)
\]  

(5)

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 \((\lambda_0)\). When the subsidy level is positive, we find

\[
\frac{\Delta \beta}{\Delta s} \bigg|_{s > 0} = 1 + \lambda_0 \lambda_m. \tag{6}
\]

We likewise see that

\[
\frac{\Delta \beta}{\Delta \lambda_0} = 1 + \mathbf{1}_{(s > 0)} \lambda_c + s \lambda_m \tag{7}
\]

which in the case of crowding out is declining in \(s\). Equations (5), (6), and (7) make it clear that when \(\lambda_c\) and \(\lambda_m\) 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 (5) or (6) is positive we say that incentives and social preferences are complements.

Using (5) and (6) we say that

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

Note that when \(\frac{\Delta \beta}{\Delta s} < 1\), 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 \(\lambda_c\) and \(\lambda_m\) or \(\lambda_0\) are zero (values are not state dependent, or they are absent). Strong crowding out holds if \(\frac{\Delta \beta}{\Delta s} < 0\) 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 \(\lambda_c\) and \(\lambda_m\) or \(\lambda_0\) zero.

The partial effect of the subsidy on the citizen’s contribution in equation (4) (that is conditional on a given level of contributions by others) is thus (for \(s > 0\))

\[
\frac{\partial a_i^{BR}}{\partial s} = \frac{1 + \lambda_0 \lambda_m \phi'' - \phi'''}{g'}. \tag{8}
\]

But the planner is interested in the effect of the subsidy on the Nash equilibrium level of contribution, and unless the benefit function \((\phi)\) is linear, the citizen’s contribution level
will depend on the contribution levels of others, as these will affect the marginal benefit of contributing. From equation (4) we know that

$$\frac{\partial a_i^{BR}}{\partial a_j} = \phi'' g'' - \phi''.$$  

(9)

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 (4) note that in order to be a mutual best response, the Nash contribution \(a^*\) must satisfy

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

(10)

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

$$n \left| \phi''(na^*) \right| - g''(a^*) < 0,$$  

(11)

a condition that ensures that the series of reciprocal effects of the citizens’ actions converges.

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.

**Assumption**

**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., \(n \phi'(na) - g'(a) > 0\), where \(a\) denotes the contribution level without subsidy.)

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

The sophisticated planner affects citizens’ contributions by selecting \(s\) to implement a Nash equilibrium given by equation (10), 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 (10) we find expressions of the derivatives of \(a^*\) with respect to \(s, \lambda_c, \lambda_m\) as follows:

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

(12)
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 (12) make clear.

To see that $\partial a^*/\partial s$ is the sum of a direct effect on the individual’s best response function and indirect effects operating via other individuals contributions, we note that

$$\frac{\partial a^*}{\partial s} = \sum_{j \neq i} \frac{\partial a^*_{BR}}{\partial a_j} \frac{\partial a^*}{\partial s} + \frac{\partial a^*_{BR}}{\partial s}.$$  

where the first term on the right hand side is the reciprocal effects of the actions of the citizens on each others actions, positive or negative depending on whether contributing is a strategic complement or substitute. From equations (8) and (9), we find

$$\frac{\partial a^*}{\partial s} = (n-1) \frac{\phi''}{g''-\phi''} \frac{\partial a^*}{\partial s} + \frac{1 + \lambda_0 \lambda_m}{g''-\phi''}$$

and rearranging gives the same expression for $\partial a^*/\partial s$ in (12). Thus the ratio of the Nash effect to the partial effect is

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

which shows that when $\phi$ is convex, there is a positive social multiplier reflecting the fact that increasing returns in the benefit function makes the citizens’ contributions strategic com-

---

**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 $\phi(\sum a_i) = \bar{\phi} \sum a_i$. 

plements, so that the partial effect of the subsidy on the individual’s contribution (holding others’ contributions constant) will be less than the total effect which includes the reciprocal indirect effects of each citizen’s contributions on the marginal benefits of other citizens’ contribution. The asymptotic stability condition (11) ensures that the social multiplier in the case of the convex net benefit function converges. From \( g'' - n\phi'' > 0 \) and \( \lambda_0\lambda_m > -1 \), an increase in the subsidy induces a higher contribution (\( \partial a^*/\partial s > 0 \)), while crowding out (negative values of \( \lambda_c \) or \( \lambda_m \)) reduces \( a^* \) (\( \partial a^*/\partial \lambda_c > 0 \), \( \partial a^*/\partial \lambda_m > 0 \)).

Note from the definition of crowding in and out, equation (4) (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 (\( \partial a_i^{BR}/\partial s < 0 \)) 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

\[
\max_{a \in [0,1], s \in [0,1]} \omega(a, s) := \phi(na) - g(a) - c(s) \\
\text{subject to } a = a^*(s, \lambda_c, \lambda_m).
\]
Figure 3: **Planner’s indifference loci** (more preferred loci are above and to the left of less preferred). The indifference map shown is computed using the following values. \( \phi(x) = 0.11x, g(x) = \frac{1}{2}x^2, c(s) = \frac{1}{2}s^2 \).

To exclude uninteresting cases we assume that \( c(s) \) is sufficiently convex so that there is a unique interior solution of (13) (See the Appendix)\(^2\).

We show indifference loci of the social planner in Figure 3.

To investigate underuse and overuse of incentives by the naive planner, as before we define the marginal rate of substitution (\( \sigma \)) and the marginal rate of transformation (\( \tau \)):

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

(14)

where the former evaluated at the citizen’s Nash response \( a^*(s) \) is just the effectiveness of the subsidy in altering Nash contributions (from (12), the slope of the implementation functions in Figure 2) and the latter 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 in Figure 3).

Then the first order condition for an interior equilibrium of the sophisticated planner’s problem equates the marginal rate of transformation with the marginal rate of substitution:

\[
\tau(a(s, \lambda_c, \lambda_m), \lambda_m) - \sigma(a^*(s, \lambda_c, \lambda_m), s) = 0.
\]

(15)

\(^2\)Notice we assume that the social planner treats the citizens’ values as a component of individual motivation, does not include them in the objective function (13). Thus, from a normative point of view we exclude from the planner’s welfare function the subjective value-based ‘satisfaction’/‘dissatisfaction’ associated with 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 in the planner’ objective function).
5. Optimal incentives with state dependent preferences

First we find the effect of categorical crowding on the optimal subsidy \( (d s^*/d \lambda_c) \). To do this, we totally differentiate (15) with respect to \( s \) and \( \lambda_c \) and find that

\[
\frac{d s^*}{d \lambda_c} = -\frac{1}{\left( \frac{\partial \tau}{\partial a} - \frac{\partial \sigma}{\partial a} \right) \frac{d a^*}{d s} - \frac{\partial \sigma}{\partial s}} \left[ \frac{\partial \tau}{\partial a} - \frac{\partial \sigma}{\partial a} \right] \frac{\partial a^*}{\partial \lambda_c}. \tag{16}
\]

Since \( c(s) \) is sufficiently convex, the second order condition for the social planner’s maximization problem is satisfied and thus the denominator of (16) (which is just the derivative of the left hand side of (15) with respect to \( s \)) is negative. Note that when categorical crowding out occurs (i.e., \( \lambda_c \) 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 \( \partial \tau/\partial a - \partial \sigma/\partial a \). In Figure 4 Panel A we consider the simplest case, in because we assume \( \phi'' = 0 = g''' \) the reduction in \( a \) does not affect the marginal rate of transformation. Then because by A2 the effect must be to reduce the marginal rate of substitution (due to diminishing marginal net returns to the project). This flattening of the planners’ indifference loci (for the given \( s = s^N \)) displace the tangency of the indifference loci and the implementation function to the right, so \( s^* > s^N \). Notice that this argument holds as long as the optimal choice in in the interior.

Concerning marginal crowding, similarly we find

\[
\frac{d s^*}{d \lambda_m} = -\frac{1}{\left( \frac{\partial \tau}{\partial a} - \frac{\partial \sigma}{\partial a} \right) \frac{d a^*}{d s} - \frac{\partial \sigma}{\partial s}} \left[ \frac{\partial \tau}{\partial \lambda_m} + \left( \frac{\partial \tau}{\partial a} - \frac{\partial \sigma}{\partial a} \right) \frac{\partial a^*}{\partial \lambda_m} \right]. \tag{17}
\]

which using using (12) and (14) can be rewritten:

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

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 bracket in (17). Note compared to the case of the categorical crowding, the marginal rate of transformation is directly affected by \( \lambda_m \) (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 stronger than the case of categorical crowding if it is to offset the direct effect of flattening of the best responses (See Figure 4 Panel B).
Figure 4: The Naive Planner’s Under-use of Incentives when Crowding out Occurs: Illustration of Proposition 1. In the case of categorical crowding out, lower values of $a$ (for a given $s > 0$ occasioned by the downward shift in the $a^*(s)$ function) reduce the marginal rate of substitution between $a$ and $s$ (“flatten” the indifference locus) so that the tangency required by the planner’s first order condition must be to the right of $s_N$. As a result the naive planner under-uses the subsidy (Panel A). To illustrate the intuition behind categorical crowding out leading to under use 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 (the right panel), either under- or over-use may occur. For under-use by the naive planner to obtain (as shown in the figure), the flattening effect of the reduction in $a$ on the planner’s indifference locus must be sufficient to offset the direct effect of marginal crowding out in reducing the marginal rate of transformation, that is slope of the Nash equilibrium best response. Panel B shows the joint effect of the reduction in the marginal rate of substitution between $a$ and $s$ (the flatter indifference locus) and the reduced marginal rate of transformation between $a$ and $s$.

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

$$\frac{ds^*}{d\lambda_c} < 0 \quad \text{if and only if} \quad \frac{\partial \sigma}{\partial a}(a^N, \lambda_m) - \frac{\partial \tau}{\partial a}(a^N, s^N) > 0$$

$$\frac{ds^*}{d\lambda_m} < 0 \quad \text{if and only if} \quad \frac{\partial \sigma}{\partial a}(a^N, \lambda_m) - \frac{\partial \tau}{\partial a}(a^N, s^N) > \frac{1}{s_N}$$

Notice that Proposition 1 provides the characterization of marginal and categorical crowding “locally”; it asserts that whenever the conditions on $\text{MRT}(\tau)$ and $\text{MRS}(\sigma)$ hold at the naive planner’s choice, a small decrease in $\lambda_c$ (or $\lambda_m$) from it 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 small degree of crowding out obtains. In Figure 5, we show categorical and marginal crowding out adopting simple functional forms (See the Appendix for the analysis).

Notice that (as in Panel C of Figure 5) the sophisticated planner adopts 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 $\lambda_m$ in Panel B, occurs because the sign of the terms in brackets in (17) changes due
Figure 5: **Optimal choices by the sophisticated planner.** Negative (positive) values of $\lambda_c$ and $\lambda_m$ imply crowding out (in). In the case of the categorical crowding out (Panel A), the optimal subsidy increases as $\lambda_c$ decreases (meaning stronger categorical crowding out) until at some critical value of $\lambda_c$, 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 the different value of parameters. In Panel B in the case of crowding out (negative values of $\lambda_m$) for any level of marginal crowding out greater than -0.8, the sophisticated planner implements a larger subsidy than the naive planner and a smaller subsidy in the case of all degrees of crowding in. In Panel C the sophisticated planner implements a smaller subsidy than the naive planner independently of whether crowding is in or out. 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.

In panel C we have chosen parameters such that the expression changes sign at $\lambda_m = 0$.

6. 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 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 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 aggregations 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 frame-
work, 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 the actor’s 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 processes by which a state
may come about often reveal important information about the intentions of others, and
they often provide 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-payoff-based 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.

**Acknowledgement.** We thank Seung-Yun Oh and [to be completed] for comments and
the Behavioral Sciences Program of the Santa Fe Institute and the U.S. National Science
Foundation for support of this project.
Appendix A. Maximization problems in Section 2 and Section 3

Here we determine the conditions under which a symmetric Nash equilibrium exists. We suppose that $s > 0$ for simplicity. Recall that from the citizen’s maximization problem (3), we have the following first order condition:

$$
\phi'(a_i^{BR} + \sum_{j \neq i} a_j) - g'(a_i^{BR}) + s + \lambda_0(1 + \lambda_c + \lambda_ms) = 0.
$$

Since we are looking for a symmetric equilibrium, we denote the equilibrium by $(a^*, a^*, \cdots, a^*)$ so the first order condition can be rewritten as

$$
\phi'(na^*) - g'(a^*) + s + \lambda_0(1 + \lambda_c + \lambda_ms) = 0.
$$

For such an $a^*$ to exists and lies in the unit interval, we need

$$
\phi'(0) - g'(0) + s + \lambda_0(1 + \lambda_c + \lambda_ms) > 0 \quad (A.1)
$$

$$
\phi'(n) - g'(1) + s + \lambda_0(1 + \lambda_c + \lambda_ms) < 0. \quad (A.2)
$$

Then when $\lambda_0\lambda_m > -1$ and $\lambda_c > -1$ hold, (A.1) is satisfied. When $g'(1)$ is sufficiently high, (A.2) is satisfied. The Nash equilibrium $a^*$ will be unique if

$$
\text{for all } a, \quad n\phi''(na) - g''(a) < 0. \quad (A.3)
$$

When $g$ is sufficiently convex, (A.3) is satisfied. In particular, when the stability condition $|n\phi''(na)| - g''(a)$ is satisfied, (A.3) is satisfied. Thus under condition (A.1) – (A.3), there exists a unique symmetric Nash equilibrium.

For the social planner’s maximization problem to yield a unique interior solution, it suffices that (i) the marginal rate of transformation is greater than the marginal rate of substitution at $s = 0$ $(\tau(a(0, \lambda_c, \lambda_m), \lambda_m) > \sigma(a(0, \lambda_c, \lambda_m), s))$ and (ii) the marginal rate of substitution is greater than the marginal rate of transformation at $s = 1$, and (iii) for a given $\lambda_c$ and $\lambda_m$ the marginal rate of substitution as a function of $s$ $(\sigma(a(s, \lambda_c, \lambda_m), s))$ increases faster than the marginal rate of transformation as a function of $s$ $(\tau(a(s, \lambda_c, \lambda_m), \lambda_m))$ as subsidy increases. For (i), since we know that $\tau > 0$, $c'(0) = 0$ ensures this holds. For (ii) and (iii), we require that $c'(1)$ is sufficiently high and $c$ is sufficiently convex.
Appendix B. Stability Condition for the Nash equilibrium

We consider the following continuous best reply dynamics (See for example Young, 1998, p.35):

\[ \frac{da_i}{dt} = a_{BR}^i(a_{-i}) - a_i, \quad i = 1, \ldots, n, \]

where \( a_{-i} := (a_1, \ldots, a_{i-1}, a_{i+1}, \ldots, a_n) \). Then it can be shown that

\[ \sum_{j \neq i} \left| \frac{\partial a_{BR}^j}{\partial a_i} \right| < 1 \quad \text{for all } i \quad (B.1) \]

implies the asymptotic stability for \((a^*, a^*, \ldots, a^*)\) (See for example the Appendix in Carpenter et al. (2009)). Then from (9) the condition (B.1) becomes

\[ (n - 1) \left| \frac{\phi''(na^*)}{g''(a^*) - \phi''(na^*)} \right| < 1 \quad \text{for all } i \quad (B.2) \]

and it is easy to see that the condition (11) implies (B.2).

Appendix C. Example in Section 4

We adopt the following simple functional forms:

\[ \phi \left( \sum a_j \right) = \bar{\phi} \sum a_j, \quad g(a_i) = \frac{\kappa}{2} a_i^2, \quad c(s) = \frac{\gamma}{2} s^2 \quad (C.1) \]

where \( \frac{\bar{\lambda}}{n-1} < \bar{\phi} < \frac{\kappa}{2} - \bar{\lambda}, \quad \kappa > 0, \quad \gamma > 0. \)

Then by the choice of \( \bar{\phi} \), the assumptions A1 is satisfied and since \( \phi'' = 0 \), A2 is satisfied. In this case we find

\[ \tau(a, \lambda_m) = \frac{1 + \lambda_0 \lambda_m}{\kappa}, \quad \sigma(a, s) = \frac{\gamma s}{n\bar{\phi} - \kappa a} \]

and

\[ \frac{\partial a^*}{\partial s} = \frac{1 + \lambda_0 \lambda_m}{\kappa}, \quad \frac{\partial \sigma}{\partial a} = \frac{\gamma s}{(n\bar{\phi} - \kappa a)^2}, \quad \frac{\partial \sigma}{\partial s} = \frac{\gamma}{n\bar{\phi} - \kappa a}, \quad \frac{\partial \tau}{\partial a} = 0. \]

From A2 we have \( n\bar{\phi} - \kappa a^* > 0 \) and thus the second order condition for the social planner’s problem \(( (\frac{\partial \sigma}{\partial a}) \frac{\partial a^*}{\partial s} - \frac{\partial \sigma}{\partial s} < 0) \) is satisfied. For pure categorical crowding out \((\lambda_m = 0)\),
there exists \( \bar{\lambda}_c \) such that

\[
\frac{ds^*}{d\lambda_c} < 0 \quad \text{for} \quad 0 > \lambda_c > \bar{\lambda}_c \quad \text{and} \quad \frac{ds^*}{d\lambda_c} = 0 \quad \text{for} \quad \lambda_c < \bar{\lambda}_c \quad \text{(Figure 5 Panel A)}.
\]

In the case of the marginal crowding out, the under-use condition in Proposition 1 becomes

\[
\frac{\partial \sigma}{\partial a}(a^*, s^*) - \frac{\partial \tau}{\partial a}(a^*, \lambda_m) = \frac{\gamma s^*}{(n \phi - \kappa a^*)^2} \kappa > \frac{1}{s^*}. \tag{C.2}
\]

Then notice that at an equilibrium, we have

\[
1 + \frac{\lambda_0 \lambda_m}{\kappa} = \tau(a^*, \lambda_m) = \sigma(a^*, s^*) = \frac{\gamma s^*}{n \phi - \kappa a^*}
\]

and from (C.2), we find

\[
\left(1 + \frac{\lambda_0 \lambda_m}{\kappa}\right)^2 = \left(\frac{\gamma s^*}{n \phi - \kappa a^*}\right)^2 > \frac{\gamma}{\kappa}
\]

Thus we see that the condition (C.2) holds if and only if \( 1 + \lambda_0 \lambda_m > \sqrt{\kappa \gamma} \). Thus for \( \bar{\lambda}_m := 1/\lambda_0(\sqrt{\kappa \gamma} - 1) \),

\[
\frac{ds^*}{d\lambda_m} < 0 \quad \text{if and only if} \quad \lambda_m > \bar{\lambda}_m \quad \text{(Figure 5 Panel B, C)}.
\]


