Building social capital: Conditional cash transfers and cooperation

Orazio Attanasio [a,*], Sandra Polania-Reyes [b], Luca Pellerano [c]

[a] *University College London, IFS, NBER, CEPR and BREAD, Gower Street, London WC1E6BT, UK*
[b] *University College London and University of Siena, Gower Street, London WC1E6BT, UK*
[c] *IFS and Oxford Policy Management, 7 Ridgmount Street, London WC1E 7AE, UK*

A R T I C L E   I N F O

A B S T R A C T

Many conditional cash transfer (CCT) programs have important social components and, therefore, can have an effect on social capital. In 2007, we conducted a field experiment with 1451 subjects in Cartagena, Colombia. We interpret the behavior in the game as a measure of what in the literature has been called social capital. We played the game in two similar and adjacent neighborhoods. The 'treatment' neighborhood, *Pozón*, had been targeted for over 2 years by a CCT program, *Familias en Acción*; the 'control' neighborhood, *Ciénaga*, had not. In 2008, with the program being implemented in both neighborhoods, we played the same public goods game, and were therefore able to implement a difference in differences strategy to estimate the impact of the CCT on our measure of social capital.

In 2007, the level of cooperation we observed in the treatment neighborhood was considerably higher than that in the control one. Although similar in many dimensions, the two groups turned out to be significantly different in some observable variables; the positive result was robust to controls for these differences. In 2008, we found that the level of cooperation was statistically identical across the two neighborhoods, and similar to the levels observed in 2007 in the treatment one. We conclude that the CCT program did improve cooperation.

In analyzing the effect of the CCT on cooperation we also look at other (individual and group) determinants of individual behavior in the game, and we compare our measure based on behavior in the game to more traditional measures of social capital used in the literature that we collected in a context-specific survey.

© 2015 Elsevier B.V. All rights reserved.

1. Introduction

Social policies may also improve economic outcomes through changes in the structures of social relationships and their forms of organization. Social capital refers to the set of resources that inheres in those relationships and the structure of such relationships (Uphoff, 1999). It can be understood in terms of social norms and networks (Putnam, 1993; Coleman, 1990) and it manifests itself in patterns of pro-social behavior (trust, reciprocity, and cooperation (Christoforou and Davis, 2014).

Conditional cash transfers (CCTs) have become one of the most popular interventions in developing countries. There is a strong line of research on the effects of CCTs that show that CCTs are successful in their goals as social assistance, raising

---

service use and improvements in target outcomes (Fiszbein et al., 2009; Adato and Hoddinott, 2010). Although CCTs' designs were meant to have a 'social capital' approach[1] and anecdotal and qualitative evidence has been found (Cohen and Franco, 2006), rigorous studies on whether CCTs affect social capital has been a prevalent interest just recently and mostly on voting behavior.[2] Although the main goals of the CCTs are not related directly to improve participation and collective action, the fact that most of CCTs have a component of new interactions among the beneficiaries within the communities would lead to build or strength social ties (Putnam, 1995) and enforce pro-social norms.

In Colombia, there is a large CCT program, called Familias en Acción (FA) that has, like many other similar programs, especially in Latin America, a strong social component. Beneficiaries meet periodically and elect a representative that coordinates a number of activities in which they are involved. It is reasonable to argue that these activities could increase cooperation, trust and, more generally, interactions among beneficiaries that could improve the social fabric of poor communities and, potentially, improve collective action and the ability to overcome social dilemmas. The main goal of this paper is to test this hypothesis exploiting quasi-experimental variation induced by the roll out of FA to urban areas.

In this study, we estimate the impact of the introduction of a conditional cash transfer program on social capital. The first challenge we had to face was to devise a measure of social capital in the context we were studying. On one hand, integrating quantitative with qualitative measures as an improved approach to measure social capital in program evaluation is not new (Adato, 2008, Casey et al., 2012) but evidence is not conclusive.[3] Adato (2000) uses qualitative data (e.g. focus groups and interviews) and finds positive link between social capital and the CCT Progresa in Mexico whereas Camacho (2014) uses survey measures only and finds no effects of the local CCT on social engagement in Peru.[4]

On the other hand, despite the complexity of conceptualization of social capital, what lies at its heart is the ability to overcome free-rider incentives in real-world situations (Coleman, 1988). These situations, that especially poor communities face, share the same game-theoretic representation. Then, as a community level attribute (Coleman, 1990), in order to build social capital, individual behavior might be subject to "free-riding" (Samuelson, 1954; Olson, 1965; Grossman and Hart, 1980; Ostrom, 1998). Evidence on social programs evaluation supports this claim (Adato et al., 2005; Fearon et al., 2009; Avdeenko and Gilligan, 2014).

For this reason, we propose a measure of social capital based on the behavior in a public goods game. The use of 'lab in the field' experiments as a method to study social preferences within a community and to measure social capital is not new.[5] In particular, public good games have been used in a variety of different situations, both urban and rural, and are one of the experimental methods that have been often used in the literature to examine behavioral motivations such cooperation in groups. Our design has been used extensively in Colombia (Cárdenas and Jaramillo, 2007; Cárdenas et al., 2013; Attanasio and Pellerano 2012) and in other countries (Cárdenas, 2003; Cárdenas et al., 2009, 2013; Barr et al., 2014). Our original contribution is to implement the game to measure the impact of a specific intervention in a quasi-experimental setting.

The Public Goods game we use to measure social capital captures the willingness to cooperate among the members of a group of 25 people by choosing simultaneously whether to allocate a token in the private account with a private benefit or to allocate the token in the group account, where the benefits of all members increases and the well-being of the entire group is improved. The sum of tokens allocated to the public good is then multiplied by a factor to model the additional value generated from the public nature of the good and then shared equally among the group members. The situation just described constitutes a typical social dilemma. There is no incentive to invest in the group account due to a higher individual payoff by investing in the private account. The dominant strategy is not to contribute at all, undermining the social outcome. However, if all in the group invest their token in the private account (assuming players maximize their own monetary payoff and rationality is common knowledge), the group will be worse-off than if all the members invest in the group account, which constitutes the social optimum. The experimental literature has extensively documented that, in similar contexts, the Nash equilibrium is not observed, either in the lab or in the field. Groups of individuals seem to be able to internalize at least in part the externality built in the game.[6] For this reason, this set up seems particularly attractive in devising a measure of 'social capital'.

---

[1] For example, Juntos (Together) in Peru, Puente (Bridge) in Chile, FA (Families in Action) in Colombia, Progresa (Progress) in Mexico, Bolsa Familia (Family pot) refer all to features of social capital: networks, 'bridging' social capital, family social capital, 'linking' social capital, cooperation (ECLAC, 2005; Arriagada and Mathivet, 2007). ECLAC (2005) present this approach in detail.

[2] Some recent evidence on the positive political impacts and voting behavior of CCT programs in Colombia (Nupia, 2012; Zárate et al., 2013), Brazil (Zucco, 2011), Romania (Pop-Eleches and Pop-Eleches, 2012), Uruguay (Manacorda et al., 2011), Philippines (Labonne, 2013) and Mexico (De la O, 2013).

[3] There is a strong line of research on the impact of "community-driven development" (CDD) programs whose main goal is to actually build social capital and seem to have no effect on the desired outcomes. Although, many of these studies suggest the existence of bias in self-reported levels of pro-sociality and social cohesion from surveys (Avdeenko and Gilligan, 2014; Mansuri and Rao 2012; Wong, 2012; Fearon et al., 2009; Casey et al., 2012), very few conduct labs in the field (Avdeenko and Gilligan, 2014; Fearon et al., 2009; Nguyen and Rieger, 2014). Fearon et al. (2009) employed a framed public goods game and found a positive impact of a CDD initiative on individuals' ability to overcome collective action problems in Liberia. Nguyen and Rieger (2014) found a positive impact on a CDD and contribution in a public goods game in Morocco.

[4] There is similar evidence from Cash Transfer (CT) programs (MacAuslan and Riemenschneider, 2011, in Malawi and Zimbabwe and Cameron and Shah, 2014, in India). However, Pellerano et al. (2014) find a positive link of a CT and engagement in reciprocal sharing activities in the community in Lesotho.

[5] Carpenter (2002), Carpenter et al. (2004), Gächter et al. (2004), Karlan (2005), Cárdenas et al. (2009), Fearon et al. (2009), Voors et al. (2012), Gilligan et al. (2013).

[6] In general, contributions to the public good are in a range of 40–60% of the group optimum (Ledyard, 1995; Croson, 1996; Camerer and Fehr, 2004; Attanasio et al., 2009). At the same time, there seems to be considerable amounts of heterogeneity in the ability different groups have to solve this kind of problems.

The possibility of cooperation within a group is determined by multiple factors such as repetition, communication, punishments or rewards and inequality in the payments. In our game, the incentives to invest in the group account are given by the specific features of the design, but also by the individual motivations concerning the group wellbeing. Individual attributes such as altruism, trust, social distance from the other members (Cárdenas, 2003), fairness (Rabin, 1993), reciprocity (Andreoni, 1988, 1995; Bowles and Gintis, 2004), a sense of affiliation as a member of a common group, or sympathy toward others in the group (Attanasio et al., 2009) determine social cohesion in a group and strengthen the ability of its members to cooperate and overcome collective action problems. In addition, community attributes that would be affected by the CCT, such as social norms and institutions, informal enforcement mechanisms, concerns for social reputation, social reciprocity (Bowles and Gintis, 2004) and group identification enforce the group interests over the individual, leading to attain a higher level of contribution and overcome the dilemma. We expect that our design will allow us to examine how cooperation has been affected by the presence of a CCT.

The first challenge we had to face is related to how we estimate the effect of the CCT on willingness to cooperate. To evaluate the impact of an intervention one would like to rely on a Randomized Control Trial. As the outcome we are interested on is the 'social capital' in a community, one would ideally randomize the intervention across a number of communities. Unfortunately, such a design was not possible. We therefore rely on evidence on individual behavior and comparing the choices of individuals living in a community with the program with that of individuals living in a community without the program. Of course, any observed difference between the 'treatment' and 'control' individuals could be driven by systematic differences at the community level in the level of social capital that are not related to the program. To circumvent this problem, therefore, we rely on a relatively standard diff in diff approach and estimate differences between the two communities we consider that are unrelated to the program by observing behavior in a period in which both communities are either both exposed or both not exposed to the program.[7] For such a strategy to be valid it is necessary that systematic differences between the two communities are constant over the period considered and determine behavior in an additive fashion. In the literature that uses diff in diff this assumption is often referred to as a common trend assumption. In what follows, we discuss its plausibility.

2. *Familias en Acción* and its expansion

Familias en Acción is a CCT that was inspired by the Mexican CCT PROGRESA and whose goal was to reduce extreme poverty in the medium term by providing resources to improve the nutritional status of poor households and in particular their children as well as school enrolment. To get access to the program's grants, beneficiary households have to comply with a number of requirements. FA has three components: a nutritional and health component aimed at households with children less than five, an education grant for children in primary school and an education grant for children in secondary school. The health and nutrition grant, roughly equal to US$25 independent of family size, is conditioned on attending regularly growth and development check-ups for children, a vaccination program and some 'classes' on hygiene, diet and contraception. The educational grants, aimed at households with children aged seven to seventeen, are conditional on enrolment and regular attendance in school. Each child in primary (secondary) school entitles the household to about US$8 (US$16) per month. Households receive a total transfer which may oscillates between 10% and 21% of the minimum wage and between 25% and 50% of the average level income of the poor (DNP, 2010; MESEP, 2012).

The program also includes an important social component, articulated around periodic meetings of beneficiaries, called *Encuentros de Cuidado* (EC) [Care follow-up meetings]. Although participation to these meetings is neither compulsory nor part of the conditions to receive the transfer, most beneficiaries (95.94% in our sample) participate in the EC where, in addition to discussing hygiene, nutrition or other health-specific issues, they have the possibility to talk about different topics or even simply chat. Beneficiaries are invited to attend the meetings, which are introduced as key for human capital investment. Conversations with program's officials and with beneficiary mothers indicate that these social aspects are indeed an important feature of the program: beneficiary mothers start new activities, get to know each other better and improve their ability to act as a group.

Additionally, the beneficiaries elect a representative, called *Madre Líder* (ML) who is in charge of communication with the local office and is also in charge of organizing the social activities and educational meetings (such as the EC) to which the beneficiaries are supposed to participate. As a consequence, the ML's often assume a prominent and visible role in the community.[8]

As the theory of social capital suggests, the fact that our CCT would require people to interact with others may create an environment for them to experience the potential of social networks to improve their wellbeing. Our CCT would affect social capital once the program is able to affect social interactions and their environment (Coleman, 1988). FA may create networks or strengthen the current ones and improve the structure of social relationships among beneficiaries (Putnam, 1995), promote leadership (Latham and Saari, 1979; Bass, 1991) and give mothers the opportunity to start working as a "social

---

[7] In our case we have one year (2007) when one community was exposed to the program and the other not and one year (2008) in which they were both exposed to the program. We therefore use 2008 to identify the systematic differences between the two communities and subtract that from the difference observed in 2007 to obtain the effect of the program.

[8] By 2007, there were 800 ML with group size ranging between 50 and 80 beneficiaries.

Table 1
Number of players and sessions.

			Individuals			Sessions[c]		
			Pozón	Ciénaga	Total	Pozón	Ciénaga	Total
Baseline	2007		342	334	676	14	14	28
Follow-up	2008	Panel	237	290	527	7	11	18[b]
		Independent cross section	404	371	775	14	15	29[a]

[a] Sessions with New recruits, i.e. with 95% or more people who were playing the game for the first time (i.e. 714 subjects).

[b] Sessions with Old recruits, i.e. with 95% or more people who were playing the game for the second time (i.e. 443 subjects)

[c] Sessions that were neither New nor Old (at 95% composition threshold) are not in the Table. There were 5 of them in Pozón and 1 in Ciénaga, corresponding to 145 subjects.

group" by perceiving a strong identification with the program (Tajfel, 1982) and their power to act 'together' (Warren, 1998). This facilitates group decision-making and increases the willingness to intervene for the common good, which eventually could lead to overcoming social dilemmas by enforcing pro-social norms (Coleman, 1990).

Empirically, we are able to distinguish between attitudinal (i.e. perception of cooperation or trust in the community) and behavioral social capital (i.e. voting behavior, participation in meetings and membership in associations as well as willingness to cooperate, social networks and social norms) (Carpenter, 2002). Our CCT may increase attitudinal social capital in some ways. First, by becoming a beneficiary a new group identity emerges which would change the perception of the community traits. Beneficiaries share the same paperwork load, health check-ups, payment logistics and the same interests. Second, the EC and beneficiaries' assemblies are a place of encounter with people that face the same needs and interests. This continued interaction among beneficiaries would create and enforce social norms so beneficiaries' perception of trust is also modified.

On the other hand, our CCT may increase behavioral social capital, first, by transforming voting behavior as confidence/perceptions on the current government or a specific political party change. Second, when social relations among members of a community are strengthened, incentives to participate in organizations or meetings in order to solve problems within the community may be stronger as well.

The program has become the flagship of the Colombian government's social policy as it targets the poorest 20% of Colombian households.[9] It started in 2002 in 627 small rural areas and in 2007 was expanded to all urban areas in order to include 1.5 million beneficiary households.[10] In 2005, the FA authorities decided to pilot the program in *Pozón*, a neighborhood in the city of Cartagena. Between 2005 and the first half of 2007, the program operated in *Pozón* but had not been implemented in other neighborhoods, despite there being other two neighborhoods (*Nelson Mandela* and *Ciénaga de la Vírgen*) identified by the FA authorities as eligible to participate in the pilot.[11]

In late August 2007, just a month after our first game, a new wave of massive enrolment to the program started in every municipality in the country, regardless of its population. The program was also rolled out in all the poorest neighborhoods of Cartagena, including *Ciénaga*.

3. Data

In this paper, we utilize two types of data. First, we analyze the behavior of individuals in the field experiment that we describe below. Second, we use extensively the survey we collected at the end of the experiment. In what follows we describe the recruitment of subjects for the game in the two neighborhoods, the protocol of the games and the data we collected.

The data for the entire set of experimental and survey data contains information on a sample of 676 and 1302 participants in the baseline and the follow up, respectively. In 2007 we conducted 28 sessions, 14 in each neighborhood. In 2008 we conducted 47 clean sessions, 21 in *Pozón* and 26 in *Ciénaga*,[12] which constitutes 775 new participants and 527 individuals that had played in 2007.[13] These figures are reported in Table 1. Overall we obtained a sample with 1.451 of individuals that did not participate in any game before.

---

[9] In Colombia, most welfare programs are targeted using the so-called SISBEN score, a poverty indicator that is updated periodically. On the basis of this score, households are assigned to one of six categories. FA targets the level 1 of SISBEN and displaced people.

[10] For evidence of success of FA on the target outcomes see Attanasio et al. (2005), Attanasio et al. (2010, 2012a, 2012b), Baez and Camacho (2011); for evidence about other outcomes such as crime and voting behavior see Camacho and Mejía (2013) and Zárate et al. (2013). Attanasio et al. (2012b) shows that these positive effects are still noticeable amongst beneficiaries 10 years after their first enrolment in the program.

[11] Ciénaga and Pozón were considered very similar to the local authorities and are adjacent. The third neighborhood (*Nelson Mandela)* is geographically distant and has received a much larger influx of displaced households from the countryside. For more information see the appendix A in the appendix.

[12] In 2008 we conducted 53 sessions, 26 in *Pozón* and 27 in *Ciénaga*. However, the complexity of the logistics led to 6 sessions in which more than 5% or less than 95% participants were new or former participants (called mixed sessions). We exclude them from the analysis.

[13] In 2008 we were faced with the issue of whether to play with the same individuals or to recruit a fresh sample of players. Both strategies had advantages and disadvantages. Playing with 'former' players would allow us to construct a longitudinal data set and potentially control for individual fixed effects. However, behavior in 2008 would be conditioned to having experienced the game within a specific group in the previous year and it would require a larger sample and more funding to take this effect into account.

3.1. Experimental set up

In our design, a Voluntary Contribution mechanism (VCM), each individual in a session of 25 participants is given one token. The player has a simple discrete choice to make: she can either invest the token in what we label a 'group' account or in a 'private' account. Tokens placed in private accounts are eventually redeemed at $5,[14] while those placed in the group account are redeemed at $10. Players who placed their tokens in the private account will be entitled to $5 and an equal share of what is in the group account. The amount in the group account is divided among all players, regardless of who contributed to it: given that there are 25 players, each would receive $0.40 for each token invested in the group account by any other member in the group. Individuals' private decisions are made simultaneously and contributions are private and anonymous, that is, players are not allowed to communicate. Those who put their tokens in the group account are only entitled to a share of what is in the group account. Note that a player contributing to the group account will get more than the $5, which is the minimum she gets if she contributes to the private account, if at least 13 players contribute to the group account.[15]

The marginal propensity to cooperate also called marginal per capita return (MPCR) in this game is one of the lowest in the literature. The goal was to mimic a measure of what is called bridging social capital, the ability to overcome social dilemmas in a very large group. Instead, we argue it matches more closely the reality we are trying to depict. Given the level of deprivation in the neighborhoods we study, the intensity of the social dilemma is arguably much higher than in most other lab, or even lab-in-the-field, studies, and a low MPCR is better fitted than a higher one.

More importantly, the low MPCR makes our results more forceful. Because the power of our analysis would have been maximized if we had had a MPCR of 50%, a low value provides evidence that if the coefficient is subject to any bias it will be downward bias. This is particularly important in light of our main goal. The abovementioned studies do not measure social capital but rather correlations between game outcomes and decisions in other contexts. Because we claim to provide a social capital measure,[16] where social dilemmas are key -and hence a low MPCR appropriate-, the low MPCR provides an additional source of validity to the measure.

The dichotomous VCM makes the game easily understood by subjects and also time effective. This is perhaps one of the reasons why only two other studies employ a dichotomous VCM game comparable we employed in Colombia: Cárdenas et al. (2013) in 6 Latin American cities, Barr et al. (2014) in Albania.

In July 2007 and July 2008 we conducted the Public Goods game described above in *Pozón* and *Ciénaga*. The VCM game we described above is played in two rounds. The subjects first decided whether to contribute to the group account individually and privately, without having the possibility of communicating with other players. After the first round (but before its results are revealed), they are given the possibility of talking about whatever they like. Then subjects had to decide again what to do with their token, individually and privately, before the results of the first round are announced. The role of communication and beliefs on individual willingness to cooperate is far beyond our research question. In this paper we only analyze willingness to cooperate in the first round.

The protocol followed in the 2 years was identical. Participants were invited to come to the local public school. After collecting their identification documents and checking their names on the recruitment lists, subjects in each session were given an identification number randomly and seated in semi-circle in a classroom where the instructions of the game were read and explained. After the participants played the second round of the game described above, we collected a network questionnaire on the existing relationships among them and a relatively short survey that gathers information on a wide range of socio-economic variables.

All recruited people were given US$1.2 as part of their show-up fee in order to induce credibility and to subsidize the transportation cost from their homes or workplace to the school we assigned for the experiments stage. Once the session ended participants were paid their earnings based on the decisions in the experiments. On average each participant earned US$9.6 (COL$16867).[17]

3.2. Sampling, recruitment and allocation into sessions

We recruited the participants to the game with the help of the local office of FA (Enlace Municipal). For the first stage of our study, in 2007 the recruitment in the treatment and control neighborhoods was slightly different. In Pozón, where the program was already operating, we could contact directly beneficiaries. The ML representatives sent an open invitation to all beneficiary mothers. In Ciénaga, instead, beneficiary mothers had not yet been registered into the program. However, since the program was already planning the expansion in, the office also had list of 'future' beneficiaries in that neighborhood.

---

[14] Units in thousands; thus $5 means COL$5000 (US$2.85, according to the official exchange rate at that date TRM: US$1=COL$1753.01 (monthly mean average for July 2008, http://www.oanda.com)).

[15] At the end of the session, only the percentage of players contributing to the group account is revealed and players are paid after they take a short socio-economic survey. Participants played the game for 2 rounds. After playing the first round, individuals are allowed to talk and communicate. Ten minutes later, there is a second round. For more details, see Appendix.

[16] In fact we claim to have a measure of the effect of the program on a measure, as opposed to a mere correlation, which arguably makes the point more relevant.

[17] FX rate US$1=COL$1,951.84 (monthly mean average for July 2007, http://www.banrep.gov.co).

These lists were used in recruiting the participants in Ciénaga. In addition, we also used an open communication strategy: a person with a loud speaker announced the invitation twice a day, 1 week before fieldwork started.[18]

In both localities, mothers were allocated into session as a first-come first served basis and there was no control over who attended which session, only whether the sessions were treatment or control and the participants dwelled in their respective neighborhood. The FA list and the open communication campaign in Ciénaga proved to be effective in identifying future FA beneficiaries, as 92.51% of participants in these sessions were subsequently selected by FA as beneficiary households in 2007.[19]

Unlike in 2007, in 2008 when we conducted the game for the second time, we could use the same recruitment strategy in both neighborhoods, as beneficiary lists were available in both neighborhoods. Besides contacting old participants, invitations were also sent to 500 randomly selected new participants from the FA beneficiaries list in each neighborhood.[20] The FA office sent the invitations through the ML to those specific households in order to attend to any of the sessions held (a span of 3 days).

Conducting lab in the field experiments in large cities provide many challenges to the experimental design in terms of costs, time, recruitment and attendance rates (Ñopo et al., 2008; Candelo and Polania-Reyes, 2008). For example, implementing sessions with 25 randomly allocated individuals was impractical and infeasible. The two neighborhoods are a 2 hour-drive apart; in order to minimize 'cross-talk' and its effects – participants talking about the experiment to future players who will participate in subsequent sessions – sessions were implemented in a four-day frame with four sessions each day in each neighborhood. For example, during the first 4 days we conducted the experiments with new participants in Pozón and the following 4 days with new participants in Ciénaga. Since the time frame was too short, we gave the beneficiaries the freedom to choose the session that bests suits them. The requirements were to be in the list of randomly chosen beneficiaries. It could have been the case that people who might be neighbors arrived together to the sessions. In fact, some invited beneficiaries arrived to the session in groups.[21]

There is the possibility of contamination among subjects of different sessions: participants to a session could talk to participants to the next session on the way out, although we put lot of effort in avoiding these contacts. This type of contamination would be the same or very similar in both treatment and control. In order to control for this possibility we included in our analysis the average outcomes in previous session and the number of sessions that had occurred before.

3.3. The post-game survey

In the survey we collected at the end of the session, there is a module on individual socio-economic characteristics (e.g. age and level of education), a module on household characteristics (e.g. income, assets, household size), a module on social capital measures (e.g. membership in associations, trust), a module about FA (e.g. date of enrolment), a module about the game (e.g. whether the participant understood the game).

3.4. Measuring social capital in a FA locality

The data we collected both during and after the games allowed us to construct several social capital measures. In what follows, we consider three different sets of measures: those derived from the behavior we observe in the games, those we can derive from the socio-economic questionnaire and those we derive from the network data we also collect.

First, from behavior in the game, we consider the propensity to contribute to the public good as a measure of attitude to cooperate in a group. An index of social capital could then be the proportion of individuals who contribute to the public good. This measure captures the ability of a group to internalize an externality and the willingness individual players have to act as a group.

Second, although the use of direct survey questions on trust has been strongly questioned as a poor indicator of social capital in comparison with experimental outcomes, in order to compare our data with the literature, we included in the postgame survey traditional measures of social capital regarding trust (such as those derived from the WVS), membership in organizations and participation in political and social processes within the community. We also considered perception of cooperative attitudes. These variables have been widely used as a proxy for social capital in the non-experimental literature.

Finally, in our survey, each player was asked about her relation with all the other players, where the options given were: (a) relative, (b) friend, (c) acquaintance or (d) unknown. In addition, we also asked questions about who would the participant consider as a leader in the community. For every session, we are able to construct a relationship matrix that describes the

---

[18] This is a method commonly used to send messages in the slums, given the lack of telephones.

[19] In particular, 82.34% of participants in these sessions were subsequently selected by FA as beneficiaries in late 2007 and 10.18% of participants belonged to a household with a FA beneficiary.

[20] Old participants were sent an invitation and were allocated in a session with a specific day and time in order to reproduce the same sessions we hold in 2007.

[21] The fact that individuals are not randomly allocated into sessions allowed us to explore the role of social networks in the effect of the program on cooperation. We were able to obtain enough variation in terms of the density and quality of the network across the neighborhoods and compared the results with those of pre-existing networks in Ciénaga 2007. We also performed a sensitivity analysis and included the network data in the regression analysis, which did not affect the results.

shape of existing networks among players. This information allows us to calculate a set of variables on session connectivity and eventually control for the strength of relationships in each group. We construct measures of how well connected the participants of each session were. The simplest measure of the "social integration" of each player within the session is the number of acquaintances, relatives and friends a participant had in her session. We also use an index of connectivity defined as the total number of connections (i.e. as an acquaintance, as a friend or as a relative) in the session and a dummy if she was considered a leader (see Appendix A).

Despite social networks being a key dimension of social capital to improve socioeconomic outcomes (Coleman, 1990; Putnam, 1995), we do not focus on our networks measures since individuals were not randomly allocated in the sessions. Instead, we focus on our behavioral measure of social capital and we do describe the other measures for illustrative purposes.

3.5. Descriptive statistics

Attanasio et al. (2009) reports descriptive statistics and the tests of significance for the difference between the two neighborhoods in 2007. These statistics are also reported in the first panel of Table 2 and give a good representation of the main characteristics of this population. Participants come from very poor families, with low levels of income and education. Although the two neighborhoods are very similar in many dimensions, and they are next to each other, important differences emerge quite clearly. Participants from *Pozón* were significantly less likely to be head of household (indicating a larger presence of partners) and have better access to public services. Moreover they are more likely to be educated. Most importantly, participants from *Pozón* seem considerably less poor than those in *Ciénaga.* They are more likely to own the house where they live, to be connected to piped water, to own several durables and other assets and less likely to have a dirt floor in the house (although there is no significant difference in the level of income or food security). They are also less likely to be unemployed. Some of the differences observed in 2007 could have been induced by the program and, given the nature of the data; it is hard to disentangle the effect that FA may have had on many of the socio economic outcomes in both neighborhoods. For instance asset tenure and monthly income may be clearly affected by the fact that participants from *Pozón* had already been received the benefits from FA.

In the second panel of Table 2 we report the differences in observable background variables between Ciénaga and Pozón in 2008. The striking feature of this second panel is that, at least in terms of observable characteristics, the differences observed in 2007 are reproduced in 2008. This is important to address the concerns about differences in selection and recruitment between the two years that we discuss below.

In Table 3, we report the average level of cooperation in our sessions in 2007 and 2008, without distinguishing neighborhoods. The data from 2008 are divided between the subjects that had already played in 2007 and those that played the game for the first time. Several considerations are in order. First, although – as mentioned – the unique Nash equilibrium of that game assuming people maximize their own short-run earnings is that individuals invest their token in the private account, many individuals do deviate from the Nash equilibrium and contribute to the public good. Despite having a very low MPC and conducting the game in an urban context, the overall level of cooperation we observe in our sample is similar to that observed in similar labs in the field. In the first round, only 14.8% and 12% of the sessions did not contribute at all. Second, there is a marked difference between the behavior of 'old' and 'new' players.

Virtually none of the 'old' players contributed to the public good and they almost uniformly played the Nash equilibrium. This result is likely to be driven by learning and by the fact that in the previous year contribution rates had been relatively low so that, even in *Pozón*, those who contributed to the public good received considerably less than the other players. This finding is also consistent with lab evidence on repeated VCM games in which, most of the times, in repeated rounds players tend to converge toward the Nash equilibrium.

In our analysis below, we present the results of the independent cross section only[22]. In Table 4, we report the descriptive statistics on several measures of 'social capital' that can be obtained from the post-game survey. These measures are divided into three groups: variables based on individual participation to civic association and neighborhood activities; variables reflecting voting behavior[23] and variables derived from answers to questions about trust and perception of cooperativeness.[24] The six columns of the table refer to the measures in Pozón, Ciénaga and the difference for the baseline and the follow-up.

---

[22] In order to avoid confounding the impact of the program with learning effects. Subjects learn from other's behavior, and update their beliefs about what other will do the next time they have the chance to play. This effect is particularly strong when the subjects know each other before entering the lab (a key issue of experiments in the field). In 2007, it is very likely that subjects met and discuss the results of the activity once they left the session. Participants in 2007 that participated again in 2008 have a different characteristic from those who face the game for the first time: they may contribute as the experimental evidence affirms that contributions decrease over time but increase in the presence of communication, they may also contribute if they contributed before.

[23] The last local elections in Cartagena (Governor, mayor, members of the *Asambleas Departamentales*, Municipal council and *Juntas Administradoras Locales*) were held three months after the inscription in the program (Oct. 2007) and eight months before the beginning of the program in *Pozón* (April 2003). Presidential elections were held sixteen months after the beginning of the program in *Pozón*, in May 2006.

[24] There is a risk that the attitudinal measures of social capital were influenced by the experimental games (Carpenter, 2002) as the players answered the questionnaire after they knew the outcome of the games and they possibly might have expressed their emotions in their answers on perception of trust and cooperation.

Table 2
Demographic characteristics of the participants – 2007 and 2008.

	Variable	2007			2008		
		Pozón	Ciénaga	Difference	Pozón	Ciénaga	Difference
General characteristics	Percentage of female participants	100	97.9	2.1**	98.8	98.1	0.7
		(0.0)	(1.1)	(1.0)	(0.7)	(0.7)	(0.9)
	Average age (years)	38.0	33.6	4.5***	36.2	36.2	0.0
		(0.6)	(0.7)	(0.9)	(0.6)	(0.4)	(0.7)
	Number of years living in the neighborhood	14.5	22.4	−7.9***	14.6	22.3	−7.7***
		(0.3)	(0.5)	(0.6)	(0.5)	(0.6)	(0.7)
	Percentage displaced	19.0	8.4	10.6***	16.5	10.3	6.1**
		(2.2)	(1.4)	(2.6)	(2.4)	(1.4)	(2.7)
	Percentage household head	21.3	30.5	−9.2**	24.3	41.3	−17.0***
		(2.3)	(3.4)	(4)	(1.7)	(3.1)	(3.4)
	Percentage single	5.6	11.7	−6.1***	10.7	12.0	−1.3
		(1.4)	(1.6)	(2.1)	(1.7)	(2.7)	(3.1)
	Percentage married of civil partnership	76.9	62.9	14.0***	77.2	67.4	9.8**
		(1.7)	(3.8)	(4.1)	(2.5)	(3.2)	(4)
Educational level (percentage)	None (level 0)	2.0	4.2	−2.1	2.3	2.7	−0.4
		(10)	(1.4)	(1.7)	(0.7)	(0.7)	(0.9)
	Primary incomplete (level 1)	18.1	18.6	−0.4	21.1	20.9	0.2
		(2.6)	(2.4)	(3.5)	(2.9)	(2.4)	(3.7)
	Primary complete (level 2)	13.5	13.8	−0.3	15.0	13.6	1.4
		(2.1)	(1.5)	(2.5)	(1.9)	(2.2)	(2.9)
	Secondary incomplete (level 3)	37.4	30.2	7.2*	35.3	30.7	4.6
		(2.5)	(3.3)	(4.1)	(2.6)	(2.2)	(3.4)
	Secondary complete (level 4)	24.0	25.4	−1.5	18.2	21.7	−3.5
		(2.9)	(2.6)	(3.8)	(2.1)	(2.5)	(3.2)
	More than secondary complete (level 5)	5.0	7.8	−2.8*	8.1	10.3	−2.2
		(1.2)	(1.2)	(1.7)	(2.6)	(1.5)	(2.9)
Dwelling characteristics	Household size	5.6	5.8	−0.2	5.6	5.8	−0.2
		(0.1)	(0.2)	(0.2)	(0.1)	(0.1)	(0.1)
	Number of people per room	3.2	2.9	0.3**	3.6	3.3	0.2*
		(0.1)	(0.1)	(0.1)	(0.1)	(0.1)	(0.1)
	Percentage with dirt floor	24.3	41.0	−16.7***	30.6	25.8	4.8
		(2.1)	(3.0)	(3.6)	(2.2)	(3.1)	(3.7)
	Percentage owning own house	82.7	58.7	24.1***	66.5	51.9	14.6***
		(2.0)	(3.0)	(3.5)	(2.2)	(2.7)	(3.4)
Public services (percentage)	Water by pipe	94.7	76.9	17.8***	91.9	83.4	8.5***
		(1.0)	(3.8)	(3.9)	(1.6)	(2.9)	(3.3)
	Sewer system	64.3	12.6	51.8***	67.3	25.5	41.8***
		(2.8)	(2.9)	(4.0)	(3.2)	(3.7)	(4.8)
	Does not have electricity	1.8	5.4	−3.6**	4.9	1.9	3.0*
		(0.7)	(1.5)	(1.6)	(1.4)	(0.7)	(1.6)
Assets (percentage)	Land phone	9.6	26.6	−17.0***	8.1	26.4	−18.3***
		(1.6)	(2.1)	(2.6)	(1.7)	(2.5)	(2.9)
	Mobile phone	86.3	67.1	19.2***	76.3	68.7	7.6*
		(2.1)	(2.3)	(3.0)	(2.2)	(3.5)	(4.1)
	Bicycle	21.9	14.4	7.6**	19.4	15.2	4.1
		(3.3)	(1.8)	(3.7)	(2.5)	(2.4)	(3.3)
	Color TV	80.7	74.6	6.2*	84.4	84.5	−0.1
		(1.5)	(3.0)	(3.3)	(1.5)	(2)	(2.5)
	Washing machine	23.7	13.8	9.9***	27.5	27.7	−0.3
		(3.1)	(1.5)	(3.4)	(2.1)	(2.9)	(3.5)
	Sound player	39.8	25.1	14.6***	35.8	26.4	9.5***
		(2.8)	(2.8)	(3.9)	(2.2)	(2.6)	(3.3)
Income variables	Percentage unemployed	2.9	10.8	−7.9***	3.8	4.9	−1.1
		(0.9)	(2.3)	(2.5)	(1)	(1.1)	(1.5)
	Percentage with access to credit	67.8	66.8	1.1	72.3	69.3	3.0
		(2.4)	(2.5)	(3.4)	(1.8)	(2)	(2.7)
	Percentage with access to formal credit	21.1	22.2	−1.1	23.7	21.2	2.5
		(1.7)	(1.9)	(2.5)	(2.1)	(2.8)	(3.4)
	Percentage with food insecurity level (high)	9.6	9.9	−0.2	7.8	9.2	−1.4
		(1.4)	(1.9)	(2.3)	(1.6)	(1.6)	(2.2)
	Per capita monthly income (US$)	25.3	22.3	3.0***	30.3	27.1	3.2*
		(0.9)	(0.6)	(1.1)	(1.6)	(1.0)	(1.9)
Observations		342	334	676	346	368	714

Robust standard errors, clustered by session.
* Significant at 10%.
** Significant at 5%.
*** Significant at 1%. For year 2008 we use Independent Cross section data.

Table 3
Behavior in the VCM – first round.

	Baseline 2007	Follow-up 2008			
		Panel[a]	Difference[b]	Independent cross section	Difference[d]
Average percentage of contributors	20.0	7.3	−8.9**	29.3	9.3*
	(3.4)	(1.8)	(3.9)	(3.7)	(4.9)
Percentage of sessions with no contribution	14.8	28	9.1	10.5	−4.3
	(6.9)	(10.9)	(13.7)	(5.8)	(9.0)
Observations	676	439	878	714	1390

[a] Old participants in sessions with Old recruits, i.e. with 95% or more people were playing the game for the second time
[b] Difference is calculated with respect to the same individuals' decisions in 2007.
[c]Sessions with New recruits, i.e. with 95% or more people were playing the game for the first time.
[d] Difference is calculated with respect to all decisions in baseline 2007.
Robust standard errors in parenthesis, clustered by session.
*Significant at 10%.
** Significant at 5%.
***Significant at 1%.

Table 4
Traditional social capital measures – independent cross section.

	Percentage		2007	2008	Difference	Standard errors
	Participation in neighborhood decisions		54.3	40.9	−13.4***	3.8
	Participation in the neighborhood meetings		61.8	47.9	−13.9***	3.0
	Membership in at least one organization		38.6	26.8	−11.9***	3.7
Civic participation	If is member of any civic association	Attendance to the meetings	38.5	26.3	−12.1***	3.7
		Decision maker	33.6	23.5	−10.1***	3.5
		Leader	18.5	9.8	−8.7***	2.1
		Supports with money or voluntary work	35.7	23	−12.7***	3.4
Voting behavior	Voted in local elections (2003 and 2007)		52.4	72.4	20.0***	3.0
	Voted in presidential elections (2006)		67	77.7	10.7**	4.4
	Trust	Most people	6.7	6.6	−0.1	1.6
		Few people	62	62	0.0	2.9
Trust and cooperation perception		None	31.3	31.4	0.1	2.8
	Perception within the community	Cooperation	26.5	32.3	5.8***	2.2
		Reciprocity	24.2	14	−10.2***	2.3
		Selfishness	49.4	53.7	4.3	2.8
Observations			676	714	1390	

Robust standard errors of the difference, clustered by session. *Significant at 10%.
** Significant at 5%.
*** Significant at 1%.

In the first panel, we report the percentage of participants actively involved in neighborhood decisions or meetings on topics related to the community, and for active members in (at least one) civic organizations, the percentage of participants that attends to the meetings, are perceive her or himself as a leader, or as a decision maker and supports the association with money or voluntary work. In the second panel, we report the percentage of individuals who voted in local and presidential elections. In the third panel, we report statistics derived from a question on trust.[25] In addition, individuals may also have different perceptions on others' social preferences. They may perceive either a reciprocal (a dummy if most people help if others help), selfish (a dummy if most people care only about themselves), or cooperative (a dummy if most people help others unconditionally) behavior in their neighbors.

Compared to what was observed in other studies (see Latorre López, 2004; Polania-Reyes, 2005) the participants reported higher levels of participation in organized groups. However, our measures of civic participation are significantly lower in the follow up. Voting behavior in 2008 is significantly higher than in 2007 for both local and presidential elections, which is consistent with some studies on voting behavior and CCT (see Zárate et al., 2013; Nupia, 2012; see footnote 3 for evidence on other countries). In the first two panels, the time trend in social capital measures is the same in both years. In the local context, voting participation is different in 2007 although we are able to argue that this measure is dependent on the time of the survey: in 2007, three months passed after the local elections and it was more than a year we ask the same question in 2008. Direct measures of trust are not significantly different across time and yield particularly a low estimate of trust. However, the sign of the difference is consistent with what we would expect: the program might have reduced the percentage of people who think no one can be trusted and the percentage of people who think that most of the people in the community are selfish.

---

[25] This question was adapted from the WVS source. We added an alternative "few people can be trusted".

Table 5
Networks characteristics across sessions – independent cross section.

	2007	2008	Difference	Standard errors
Average number of friends	1.0	1.5	0.4[*]	0.2
Average number of acquaintances	7.7	0.4	−7.3[***]	0.6
Average number of connections	8.9	2.0	−6.9[***]	0.7
Friendship density[a]	0.1	0.1	0.1[***]	0.0
Acquaintanceship density	0.4	0.0	−0.3[***]	0.0
Connections density	0.4	0.2	−0.3[***]	0.0
Percentage of players identified as leader	33.4	18.1	−15.4[***]	3.3
Observations	676	714	1390	

[a] Network density is calculated as the number of connections/(number of connected people)*(number connected people − 1). For example, Friends density is calculated as the number of friends/(number of people with friends)*(number of people with friends − 1). Robust standard errors of the difference clustered by session.

[*] Significant at 10%.
[**]Significant at 5%.
[***] Significant at 1%.

In Table 5 we present the average number of friends, acquaintances and connections (the sum of relatives, friends and acquaintances) each participant reports in the session where she is playing. We also report a feature of the network in the session, the friendship, acquaintanceship and connectivity densities, measured as the ratio of the total number of identified specific links in the session and the total possible number of specific links among connected people, i.e. those individuals that are identified as an acquaintance at least once by other player. In addition, we present a measure of leadership as the percentage of players identified as leader in each session, at least by one different player in the session.

Participants in Pozón in 2007 attended sessions with higher levels of connectivity but these differences are gone in the follow up. We are not able to explain rigorously these differences as we may find that the program balanced the difference between networks in the sessions in the follow up. We may also say that the recruitment process in 2007 facilitated participants to attend in groups of people with pre-existing links. For example, in Ciénaga in 2007, participants had fewer friends than in Pozón 2007 and 2008 but in 2008 participants in Ciénaga attended the sessions with more friends.

We also find that the percentages of players identified as leaders in the session is significantly lower in the follow up, suggesting that a certain fragmentation affects the patterns of leadership consolidation in the social context of the study. Interestingly, the rate of reported leaders in baseline and follow up is significantly higher than the proportion of ML (participants who declared to have been elected FA beneficiary representatives) in the cross-section sample (9.0% and 5.1%, respectively).[26]

Surveys record stated preferences while experiment outcomes provide revealed preferences. The experimental literature on economic experiments and surveys does not find a consistent answer on whether these two are positively correlated.[27] On one hand Gächter et al. (2004) find a positive correlation between stated and revealed preferences. On the other hand Capra et al. (2008) and Cárdenas et al. (2013) find no evidence of correlation between participation in a charity and contribution in a public goods game.[28]

4. Identification issues and empirical specification

To establish the relationship between the expansion of the CCT program and social capital we compare the behavior in the game of individuals who had been exposed to the program to those who had not. In July 2007, the program had been operating in the *Pozón* for over a year, while it had not been operating in *Ciénaga*. It should be stressed that the expansion to *Ciénaga* had already been decided so that the FA office could provide us with lists of potential beneficiaries in *Ciénaga*. In July 2008, we conducted the same game in both neighborhoods, when the program was operating in both. In addition we gathered information on individual and household level variables, as well as on the links (relatives and friends) among the participants to the sessions. By the time we conducted the follow up experiments, *Ciénaga* had been already targeted for over 12 months.

The expansion to *Ciénaga* allows us to implement a difference in differences (DiD) strategy to control for the possibility that the two neighborhoods differ in some unobserved dimension, provided that such 'unobservable' differences do not change over time. Although this is an untestable assumption we show below that most structural individual and household characteristics were stable within each neighborhood during the evaluation time.

---

[26] We find that 84.2% (for baseline) and 46.2% (for follow up) among those identified as leaders in the session are MLs.

[27] This could be explained with what Carpenter (2002) refers to as the idealized persona bias, i.e. a desire to appear as the person one would like to be rather than what they are, and/or the surveyor effect, i.e. the desire to answer in accordance to what one thinks the surveyor would like to hear or see.

[28] Our study is not exceptional, we present a comparison between behavior in the game and the traditional measures of social capital with traditional measures exhibit a weak correlation with such behavior. In fact, the correlations we find are not stable over time, which reinforces the concern of using these measures in the analysis of social capital.

The first identification issue raised by our study is the lack of randomization in the allocation of the program across neighborhoods. Based on conversation with program Administrators, it was felt that the initial allocation of FA was not driven by specific pre-existing differences between neighborhoods.

Although the availability of information on a wide range of pre-intervention variables allows us to control for the possibility that the two neighborhoods differed in the level of social capital for reasons different from the operation of the program, it is possible that there are unobserved differences that we could not control for. This motivated us to perform a follow up study a year after the first stage of the study. As discussed in the introduction, in 2008 both neighborhoods were exposed to the program so that, under some assumptions, it is possible to use the data from our second experiment to estimate these unobserved differences and correct the estimates derived from considering the difference between treatment and control in 2007.

The approach we take is essentially a DiD one. The only difference relative to a standard situation where at baseline both treatment and control are not exposed to the program while at follow up only the treatment are, is that in our case only one neighborhood was exposed at baseline, while both were exposed to the treatment at follow up. Unlike in the standard case where unobserved differences between the two groups are estimated using baseline data, we use the follow up data to do the same thing.

For our approach to be valid we need a number of assumptions. First, we need to assume that, conditional on observed variables, unobserved differences between the two neighborhoods are constant over time and enter additively. This is what is usually referred to as the common trend hypothesis. Unfortunately, it is not possible to test this assumption because of the lack of retrospective information at the individual or neighborhood level. However, we discuss below why we think this is a reasonable assumption to make on the basis of the available descriptive statistics.

Second, we need to assume that, after a few months of exposure, the impact of the intervention on social capital does not change with exposure. That is, if the program affects social capital, it does so by moving, after a few months, the social capital to a new level. This assumption is necessary to allow us to identify the unobserved differences in social capital between the two neighborhoods using the 2008 data when the program had operated for over 2 years in Pozón and only for just over 11 months in Ciénaga. Such a hypothesis is plausible as intensive interaction among beneficiaries starts early in the process. We assume that the interactions between beneficiaries in the first year of exposure to the program are sufficient to bring the level of social capital at a new equilibrium level that does not change after this first initial period.

To improve efficiency of our analysis, we can also control for a number of variables that are unaffected by the program. The common trend assumption we make to use the DiD approach, has to hold conditional on these variables.

In order to model the probability to contribute to the public good, we run a simple probit regression, where a single value index Y_{ist} is modeled as follows:

$$Y_{ist} = \alpha + \beta X_{ist} + \gamma D_{i07} + \delta T_{i07} + \lambda \left(1 - D_{i07} * T_{i07}\right) + \theta S_{st} + \nu_{st} + \varepsilon_{ist} \tag{1}$$

where Y_{ist} is the individual decision to contribute to the group account in session s at time t; D_{i07} is a dummy for Ciénaga (which is the *control* group in 2007); T_{i07} is a dummy for the year 2007, when Ciénaga was not exposed to the program; X_{ist} are individual observable characteristics, S_{st} are session level characteristics and ε_{ist} are i.i.d. Gaussian distributed error term with mean zero and variance $\sigma_{\varepsilon}^2 = 1$, independently of ν_{st}, which are i.i.d., $N\left(0, \sigma_{\nu}^2\right)$. The parameter λ in the expression above captures the effect of the program on the single value index Y_{ist} and determines the impact of the program on the probability to contribute to the public good. The time dummy captures common trend effects by exploiting the difference in contribution rates between 2007 and 2008 in *Pozón* (assuming that any FA impact had reached a stable level already in 2007); the neighborhood dummy captures fixed effects by exploiting the difference in contribution rates between *Pozón* and Ciénaga in 2008 (assuming that any FA impact had reached a stable level in both neighborhoods then).

Another potentially important issue is related to our sampling method. As we discussed above, due to logistical and data availability constraints, the recruitment approach was different across in 2007 and 2008 and, in 2007 was different between treatment and control neighborhoods. In both neighborhoods and both data collection stages the sample of participants in session includes individuals that self-selected for participation into the experiment by responding to an invitation. Hence the study is not based on a pure random sample and results might be non-representative of the whole beneficiary population. However the design was established to ensure that the self-selection bias affected the treatment and control group in a similar way at each stage of the study. Table 2 above constitutes an important and reassuring piece of evidence in this respect. As we mentioned, whilst the two neighborhoods appear to be different in a number of observable dimensions, these dimensions are virtually identical over time. If the selection process had been substantially different in Ciénaga in 2007 and 2008 and had been the same in Pozón, then this would have resulted in the observed differences between the two samples to change over time.

We conclude this discussion with a consideration about statistical inferences. Observations within a small neighborhood could be correlated because neighbors share common characteristics besides the assignment into treatment and control. However, the DiD approach does take into account the presence of a systematic difference between the two subsamples. As it turns out, for many outcomes, the intra-class correlation coefficient within neighborhoods is relatively low at 0.010.[29] On

---

[29] Individuals within neighborhoods are no more similar than individuals in different neighborhoods (in part for our sample size), and we effectively assigned 1390 individuals to treatment or control. We can reject that we have only two independent observations.

Table 6
Program impact on cooperation – first round.

Independent Variables	I	II	III	IV
Dependent variable: 1 if player contributed to the group account in the first round				
Dummy *Ciénaga*	0.105*	0.084	0.048	0.064
	(0.060)	(0.057)	(0.062)	(0.052)
Dummy 2007	0.089*	0.090*	0.064*	0.096***
	(0.050)	(0.049)	(0.037)	(0.033)
Program impact	0.306***	0.293***	0.241***	0.233***
	(0.044)	(0.046)	(0.043)	(0.034)
Basic controls – participant characteristics	No	Yes	Yes	Yes
Session characteristics	No	No	Yes	Yes
Session composition (info within session)	No	No	No	Yes
Observations	1390	1384	1384	1384

Marginal effects reported. Robust standard errors, clustered by session, in brackets after probit estimation.

* Significant at 10%.
** Significant at 5%.
*** Significant at 1%.

the other hand, the behavior of participants at a given session is affected by many common factors, ranging from the identity of the game master, to the time of day to the general atmosphere in the session. Therefore we account for associations among observations within sessions (clusters) and cluster our standard errors at the session level.

5. Do CCTs build social capital?

We identify the effects of FA on our experimental measure of social capital by comparing first the contribution rates to the group account in our VCM in the two neighborhoods in 2007 and 2008. Then, we analyze the impact of the program on different and more traditional measures of social capital (Chattopadhyay and Duflo, 2004; Jones and Olken, 2005).

5.1. Identifying the effect of FA on the willingness to contribute to a public good

In Table 6, we pool together the 2007 and 2008 data to identify the effect of FA on the probability to contribute to the group account, after controlling for a variety of observable variables (the complete set of results is presented in Table 8), using the probit model in Eq. (1) in all four columns. In Column I we do not add any controls. In column II we control only for individual characteristics such as the age and education level of the participant. In Column III we control for session characteristics. An issue that worried us considerably on the field was the possibility that individuals who played in early sessions would 'contaminate' other individuals that were about to enter subsequent session by talking to them and commenting on the game. While we tried to avoid these contacts as much as possible and instructed the subjects not to talk to subsequent players, some contacts were unavoidable. For this reason, we control for the sequence order in which a particular session is played in a day and in the full sequence of sessions and also for results in previous sessions.[30] We do find these session effects to be significant. However, their introduction does not change the size of the marginal impact, partly because these effects were present in both treatment and control sessions.

Under the assumptions discussed above, the coefficient λ in Eq. (1) can be interpreted as the impact of the program and its estimates are reported in the third row of the table. The impact is large and significant at the 1% level.

The positive coefficient on the 2007 dummy indicates (under the maintained assumptions) the presence of a negative common trend in our measure of social capital. Finally, the positive coefficient on the Ciénaga dummy indicates that unobserved differences in social capital (not related to the presence of the program) are higher in Ciénaga than in Pozón.

The coefficients in the first column are the results of mean participation rates that are higher in Pozón than in Ciénaga (when the result was operating in both neighborhoods) and they flip in 2008 when the program was operating in both. Therefore the bias in estimating the impact of the program using only the data from 2007, as in Attanasio et al. (2009) would be negative. Notice, however, that the coefficient on the Ciénaga dummy becomes statistically insignificant once we control for observable variables. Notice also that if one thinks that the impact of the program takes more than 12 months to develop and that in 2008 its impact was stronger in Pozón than Ciénaga because the program had operated longer in the former than in the latter, then the estimates we present in Table 6 would be biased downwards.

In the specifications in Column III and IV, the point estimates of the impact of the program are now lower, but they are still significantly different from zero at the 5% level. The main message of the table, therefore, is that FA had a strong impact on social capital; at least as measured by the behavior in the game we played in Cartagena. This result seems to be robust to the presence of systematic differences between the two neighborhoods we have compared, both in terms of observable

[30] Particularly, in order to control for contiguous sessions correlation, we include in specification (IV): a) a dummy for the first session each day; b) a variable capturing the deviation from the neighborhood mean of the average contribution to the public account in the previous 2 sessions, and c) the order (number) in which the session came about in the full sequence of sessions within the neighborhood.

Table 7
Program impact on traditional measures of social capital.

Probit marginal effect of the program impact[a]			I	II
	Membership in at least one organization		−0.070	−0.116*
			(0.071)	(0.069)
Civic Participation	Participation in neighborhood decisions		0.236***	0.221***
			(0.049)	(0.054)
	Participation in the neighborhood meetings		0.158***	0.111**
			(0.046)	(0.050)
	Vote Local Elections (2003 and 2007)		0.159***	0.134**
			(0.055)	(0.058)
Voting behavior	Vote Presidential Elections (2006)		0.171***	0.095
			(0.061)	(0.062)
	Trust	Most people	0.005	0.003
			(0.029)	(0.027)
		Few people	0.127**	0.116*
			(0.056)	(0.062)
		None	−0.136***	−0.122**
Trust and cooperation			(0.051)	(0.057)
perception	Perception within the community	Cooperation	0.082**	0.099**
			(0.040)	(0.043)
		Reciprocity	0.013	0.009
			(0.047)	(0.045)
		Selfishness	−0.104**	−0.118**
			(0.050)	(0.052)
	Number of player's relatives		0.034	−0.185***
			(0.023)	(0.049)
	Number of player's friends		1.255***	0.671*
Individual social			(0.139)	(0.380)
integration	Number of player's acquaintances		1.097***	2.350**
			(0.347)	(1.026)
	Number of player's connections		2.386***	2.836***
			(0.352)	(1.028)
Player is identified as			0.136***	0.122***
leader in the session			(0.041)	(0.044)
Basic controls – Participant's characteristics	No			Yes
Session characteristics	No			No
Session composition (info within session)	No			No
Observations	1390			1384

Coefficients for the marginal effects in probit specifications I and II, the dependent variable being the one listed on the left hand side. Robust standard errors, clustered by session, in brackets after probit estimation.

[a] Coefficients for the measures of individual social integration are OLS estimates.

* Significant at 10%.
** Significant at 5%.
*** Significant at 1%.

and unobservable variables. Given the magnitude and robustness of the estimate, it doesn't seem plausible that this result is driven exclusively by differential trends in unobservable across Ciénaga and Pozón.

5.2. Impact on traditional measures

We now analyze the impact of the program on the measures of social capital we summarized in Tables 4 and 5. We implement Eq. (1) to model the probability that an individual engages in a 'pro-social' activity (e.g. participation in neighborhood decisions) with a probit regression[31] in which we include the same individual control variables that we considered in the range of specifications for the decision to contribute in the game in Table 6. Table 7 reports the impact of the program (the coefficient λ) on each of these measures of social capital.

A few social capital measures are statistically significant across both specifications. The clearest effect is on 'individual social integration' variables as well as on 'voting behavior'.

The effect on 'civic participation' is still positive but mixed: it is still positive except for participation in at least one civic organization (where coefficients are negative but not significant, consistent with Pellerano (2008) who, considering the rural program, observed that participation into the EC seemed to substitute for participation in other organizations).

Finally there is a clear effect on perceptions about trust and cooperation. Whilst the program does not seem to create positive perceptions about other people's attitudes toward trust and cooperation, it does seem to reduce the negative ones. This is a noteworthy result that sheds light on the manner in which the CCT can affect beliefs.

---

[31] For the measures on Individual social integration we use a linear regression model.

Overall, Table 7 depicts a picture that complements the one arising from Table 6 in terms of impact of the CCT program.[32]

6. The determinants of social capital

The evidence we have presented so far seems to indicate that FA had a strong impact on different dimensions of social capital. In the process of getting this result, in Section 5, we have controlled for several variables, at the individual, household and session level, to ensure that the results we obtained could not be explained by differences in these variables between the two neighborhoods considered. The coefficients on these variables are interesting on their own, as they are informative about the main determinants of cooperative behavior. In this section, we discuss the coefficients we obtained on the various variables we considered as determinants of the behavior in the game.

6.1. Individual and session variables effects on social capital

In the first panel of Table 8, we report the coefficients on the individual and household level variables in the regressions that were used to compute the impact of the program as reported in Table 6. Age is an important determinant of cooperation: older players seem to be more likely to contribute to the group account. The few men who participated (mainly in 2007) in the game were less likely to cooperate. As for household level variables, we find that players with a large household size tend to cooperate more. Interestingly, we do not find any education effects and very limited effects of other household level variables, including various wealth indicators.

Several of the session level variables are significant determinants of cooperation. Whilst men are less likely to contribute, the presence of a man in the session seems to induce cooperation. Players in the first session of the day were much more likely to cooperate, while the total number of players in a session had a negative effect on cooperation. Finally, when we consider (in Column IV) another set of variables controlling for the *composition* of the sessions, we find that participants in a session with a high percentage of the highest level of education (i.e. secondary education) within the session and a session socially distant with respect to other sessions in the neighborhood (in terms of the level of education of its participants) will be more willing to contribute to the public good.[33]

We also note that, although the three game administrators (who were rotated across the sessions in both neighborhoods) were instructed to read exactly the same protocol and their behavior in the sessions differed (necessarily) only when answering the questions posed by the participants, we find strong 'administrator' effects.

6.2. Leadership and social capital

An important feature of the FA program is the importance that the role of leaders acquires in the community activities and therefore their role in strengthening social ties with the communities and becoming channels of wellbeing improvement (Latham and Saari, 1979; Bass, 1991).[34] In this subsection, we investigate whether the presence of a ML in the session affected individual decisions. Table 9 presents the results we obtain if we consider, among the determinants of individual behavior, the presence of a ML among the participants. Cooperation is positively associated with the presence of at least one ML in the session. The marginal effect is large at around 13.2%, when we control for session characteristics and composition (specifications VII and VIII), and above 16% when we do not control for session these variables (see specifications V and VI). The Program impact has a lower coefficient but still positive and significant. We can propose that one of the channels through which cooperation succeeds in the community is their leader. The results we obtain on the other control variables are not different from those presented in Table 8 and are reported, for completeness, in Appendix C in Table C2.

In addition, being an ML may be a determinant of the attitudes to cooperate, not only to the beneficiary that was elected as a ML but also to those who may be considered leaders by the group despite of not having an official title. In Table C3 in the appendix we present the results when we include whether the individual is considered as a Leader in the group and the percentage of participants who consider this particular individual as one of their leaders in the group.

---

[32] The question arises as to whether Table 7 says the effect of the CCT is indirect, via the variables captured by the survey, or direct. From Table A2 we observe there is no significant link between the survey variables and the behavioral one, which provides some external validity to the claim that the program has an impact on cooperation. Having discussed the plausibility for lack of correlation between stated and revealed preferences, this should not raise a concern.

[33] Despite considering our network measures as measures of social capital instead of variables that explain cooperation, we explored the role of preexisting family and friend networks among the beneficiaries in the decision to contribute. Among all the specifications we attempt by adding different combinations of the network measures to specification (IV) in Table 6, only found a significant association between Friendship density of the session and the decision to contribute (i.e. marginal effect of 0.538, significant at 1%).

[34] Brune and Bossert (2009) find a positive effect of an intervention based on leadership development on a wide array of social capital measures from the World Bank's Social Capital Assessment Tool in Nicaragua. Macours and Vakis (2014) found that social interactions with nearby leaders positively affected human capital and productive investments in a CCT in Nicaragua.

Table 8
Controls for willingness to cooperate in first round (independent cross section).

Dependent variable: 1 if player contributed to the group account in the first round		II	III	IV
	1 if the player is a woman	0.118[**]	0.156[***]	0.158[***]
		(0.058)	(0.030)	(0.028)
	If the player has a partner	−0.002	0.006	0.008
		(0.036)	(0.033)	(0.033)
	Level of education (0–5)	−0.001	−0.007	−0.005
		(0.010)	(0.009)	(0.009)
	Household size	0.016[***]	0.012[**]	0.013[**]
		(0.006)	(0.006)	(0.006)
	Ground floor (house)	−0.045[*]	−0.031	−0.030
		(0.025)	(0.023)	(0.023)
	Number of years living in the neighborhood	0.001	0.001	0.001
		(0.001)	(0.001)	(0.001)
	1 if the player's home has water pipe access	0.043	0.038	0.044
		(0.035)	(0.032)	(0.031)
	1 if the player's home has sewage	−0.039	−0.031	−0.024
		(0.029)	(0.028)	(0.027)
	1 if the player's home has no electricity	0.128[*]	0.086	0.096
		(0.072)	(0.071)	(0.072)
Basic controls – participant characteristics	1 if the player is the head of household	0.019	0.022	0.024
		(0.039)	(0.039)	(0.040)
	Age	0.005[***]	0.005[***]	0.005[***]
		(0.001)	(0.001)	(0.001)
	1 if the player is unemployed	−0.017	0.007	0.008
		(0.057)	(0.058)	(0.058)
	1 if the player has her own housing	−0.041	−0.049[*]	−0.048[*]
		(0.027)	(0.025)	(0.025)
	1 if player belongs to the fifth quintile per capita income	0.047	0.036	0.041
		(0.031)	(0.030)	(0.030)
	1 if the player has a phone	−0.001	−0.001	−0.005
		(0.038)	(0.038)	(0.038)
	Number of rooms in the house	−0.018	−0.011	−0.010
		(0.014)	(0.014)	(0.015)
	1 if the player is displaced (self-declared)	0.025	0.015	0.013
		(0.032)	(0.032)	(0.032)
	1 if the player declares she understood everything	−0.048	−0.043	−0.044
		(0.029)	(0.030)	(0.030)
	1 if the household receives support from another institution	−0.063[**]	−0.054[**]	−0.052[*]
		(0.029)	(0.027)	(0.027)
	Number of players in session		−0.030[***]	−0.037[***]
			(0.009)	(0.008)
	1 if there is at least one man in the session		0.166[***]	0.201[***]
			(0.043)	(0.044)
	Experimenter n°2 (female) in 2007		−0.049	−0.071[*]
			(0.052)	(0.043)
	Experimenter n°3 (male) in 2007		0.121[**]	0.135[***]
			(0.047)	(0.043)
Session characteristics	Experimenter n°2 (female) in 2008		0.099[*]	0.096[**]
			(0.058)	(0.049)
	First session in the day		0.080[**]	0.095[***]
			(0.040)	(0.035)
	Behavior in the 1st round of the last two sessions [a]		−0.316[*]	−0.226[*]
			(0.162)	(0.124)
	Number of sessions held before (t)		−0.020[***]	−0.019[***]
			(0.003)	(0.003)
	Percentage of participants with less than secondary complete			0.370[***]
				(0.109)
Session composition	Mean absolute deviation - level of education in the neighborhood			0.198[*]
				(0.103)
Observations		1384	1384	1384

Marginal Probit for II–IV. Marginal effects reported. Robust standard errors, clustered by session.

[a] Calculated as the Mean of previous 2 Sessions' Deviation from the average contribution in the neighborhood.

[*] Significant at 10%.

[**] Significant at 5%.

[***] Significant at 1%.

Table 9
Role of the leader in the first round.

Independent variables	V	VI	VII	VIII
Dependent variable: 1 if player contributed to the group account in the first round				
Dummy *Ciénaga*	0.103*	0.083	−0.003	0.028
	(0.056)	(0.051)	(0.042)	(0.039)
Dummy 2007	0.056	0.059	0.060*	0.091***
	(0.050)	(0.048)	(0.036)	(0.031)
Program impact	−0.202***	−0.188***	−0.107*	−0.121**
	(0.070)	(0.070)	(0.065)	(0.056)
1 if there is at least a ML in the session	0.165***	0.166***	0.144***	0.132***
	(0.062)	(0.058)	(0.040)	(0.041)
Basic controls – participant characteristics	No	Yes	Yes	Yes
Session characteristics	No	No	Yes	Yes
Session composition (information within session)	No	No	No	Yes
Observations	1.390	1384	1384	1384

Marginal probit. Marginal effects reported. Robust standard errors, clustered by session, in brackets after probit estimation. We do not use the dummy indicating whether participant is a ML: it is collinear with program impact and the coordinator 2 in 2007. We would have to exclude the data from *Ciénaga* in 2007 as the program was not operating there and, as a consequence, there were no ML in those sessions and the DID would not be feasible.

* Significant at 10%.
** Significant at 5%.
*** Significant at 1%.

7. Conclusions

Our main purpose is to examine, by combining survey and experimental methods, the effect of a social intervention on individual and group behavior and their ability to obtain better social and economic outcomes. There are two main difficulties in testing rigorously the hypothesis that CCTs increase social capital. The first lies in the difficulty in obtaining a quantitative measure of social capital. The second is the standard evaluation problem of observing the counterfactual to a given intervention.

We addressed the first difficulty implementing a public goods game. Our main measure of social capital is based on the behavior during this game, whose rewards, for the individuals in our sample, are salient. Attanasio et al. (2009) started our analysis by comparing the level of cooperation in two similar neighborhoods in 2007 when the CCT program we have been studying was operating only in one of them and show that the level of cooperation we observe in the 'treatment' community is considerably higher than in the 'control' community. In 2008, after the program had been expanded to the control community, we retake the measurements taken in 2007 and find no strong difference between cooperation levels in the two neighborhoods. By using a difference in difference approach, we control for unobserved differences in the social fabric of the two communities and find evidence to support the hypothesis that the introduction of the CCT we have been studying did indeed increase social capital.

The other focus of the paper is on how to measure social capital. While we implemented a VCM game and used observed individual behavior in the game as our main measure, we also collected information on self-reported behavior along a number of dimensions that have been used in the literature as measures of social capital. We show that while these measures roughly correlate with ours, they are also affected by measurement error, which makes the identification of the impacts of CCTs on social capital harder to detect. Despite this, with our DiD approach we found a positive effect of the CCT on participation in neighborhood decisions, participation in the neighborhood meetings and a negative effect of the CCT on the perception that "it is not possible to trust in anybody" and in the perception that "in the community people only care about themselves". We only found a weak effect of the CCT on the number of connections subjects have within the session. We also explore the role of preexisting family and friend networks among the beneficiaries in an alternative effect of a CCT. Although it has been documented how pre-existent networks may determine the effect of a CCT program on the main outcome (Angelucci and De Giorgi, 2009; Angelucci et al., 2010; Macours and Vakis, 2014), it has not been the case with the effect of our particular CCT.

It must be noted that the external validity of our results might be limited by some of the sample selection issues we discussed above: our sample was made of self-selected individuals that responded to an invitation to participate to game sessions and the formation of groups was also left to individual self-selection into sessions depending on the timing of arrival at the game site. Despite these caveats, and whilst results are not representative to the whole population of beneficiaries of FA, they provide compelling evidence about the potential influence of CCTs on social capital formation and the mechanisms through which such effect is likely to realize.

The effect that CCT programs such as FA might have on social capital, probably because of the many social activities linked to the program, has long been debated. Some policy makers even hypothesized that the program might have a negative impact on social capital creating division between beneficiaries and (marginal) non-beneficiaries. Our evidence represents one of the first quantitative assessments of the impact of a CCT on social capital and provides strong support for the hypothesis that social capital is improved. The policy implications of these findings are obvious. This impact should be

added to the benefits that conditional cash transfers could deliver. Of course, one has to be careful and understand that the manner in which a CCT is delivered might be key to the social capital impacts we see.

Acknowledgements

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.jebo.2015.04.004.

References

Adato, M., 2000. Final report: The impact of Progresa on community social relationships. International Food Policy Research Institute, Washington, DC, Available from: http://evaluacion.oportunidades.gob.mx:8010/es/wersd53465sdg1/docs/2000/ifpri_2000_community_level_impacts.pdf

Adato, M., 2008. Integrating Survey and Ethnographic Methods to Evaluate Conditional Cash Transfer Programs, vol. 810. International Food Policy Research Institute.

Adato, M., Hoddinott, J., Haddad, L.J., 2005. Power, Politics, and Performance: Community Participation in South African Public Works Programs, vol. 143. International Food Policy Research Institute.

Adato, M., Hoddinott, J., 2010. Conditional Cash Transfers in Latin America. International Food Policy Research Institute, http://www.ifpri.org/sites/default/files/publications/conditional_cash_transfers_in_latin_america.pdf

Andreoni, J., 1995. Cooperation in public goods experiments: kindness or confusion. Am. Econ. Rev. 85 (4), 891–904.

Andreoni, J., 1988. Why free ride? Strategies and learning in public good experiments. J. Pub. Econ. 37, 291–304.

Angelucci, M., De Giorgi, G., 2009. Indirect effects of an aid program: how do cash transfers affect ineligibles' consumption? Am. Econ. Rev., 486–508.

Angelucci, M., De Giorgi, G., Rangel, M.A., Rasul, I., 2010. Family networks and school enrolment: evidence from a randomized social experiment. J. Pub. Econ. 94 (3), 197–221.

Arriagada, I., Mathivet, C., 2007. Los programas de alivio a la pobreza Puente y Oportunidades. Una mirada desde los actores. Polít. Soc. Ser. 134. CEPAL, Santiago de Chile, 40 pp.

Attanasio, O., Fitzsimon, E., Mesnard, A., Vera-Hernandez, M., 2005. How effective are conditional cash transfers? Evidence from Colombia. In: IFS Briefing Note.

Attanasio, O., Fitzsimons, E., Gomez, A., Gutiérrez, M.I., Meghir, C., Mesnard, A., 2010. Children's schooling and work in the presence of a conditional cash transfer program in rural Colombia. Econ. Dev. Cultural Change 58 (January (2)), 181–210.

Attanasio, O., Battistin, E., Mesnard, A., 2012a. Food and cash transfers: evidence from Colombia. Econ. J. R. Econ. Soc. 122 (559), 92–124, 03.

Attanasio, O., Pellerano, L., 2012. Impactos de largo plazo del programa familias en acción en municipios de menos de 100 mil habitantes en los aspectos claves del desarrollo del capital humano Final Report. DNP. Bogota, https://sinergia.dnp.gov.co/Sinergia/Archivos/92646308-df90-4107-a074-8bf62ea9e862/Eval_Familias%20Acci%C3%B3n%20Largo%20plazo.pdf

Attanasio, O., Barr, A., Cardenas, J.-C., Garance, G., Costas, M., 2012b. Risk Pooling, risk preferences, and social networks. Am. Econ. J. Appl. Econ. 4 (2), 134–167.

Attanasio, O., Pellerano, L., Polanía-Reyes, S., 2009. Building trust? Conditional cash transfer programs and social capital. Fiscal Stud. 30 (2), 139–177.

Avdeenko, A., Gilligan, M.J., 2014. International interventions to build social capital: evidence from a field experiment in Sudan.

Baez, J.E., Camacho, A., 2011. Assessing the long-term effects of conditional cash transfers on human capital: Evidence from Colombia. World Bank Policy Research Working Paper Series, vol. 2011.

Barr, A., Packard, T., Serra, D., 2014. Participatory accountability and collective action: experimental evidence from Albania. Eur. Econ. Rev. 68, 250–269.

Bass, B.M., 1991. From transactional to transformational leadership: learning to share the vision. Organ. Dyn. 18 (3), 19–31.

Bowles, S., Gintis, H., 2004. The evolution of strong reciprocity: cooperation in a heterogeneous population. Theor. Popul. Biol. 65, 17–28.

Brune, N.E., Bossert, T., 2009. Building social capital in post-conflict communities: evidence from Nicaragua. Soc. Sci. Med. 68 (5), 885–893.

Camacho, L.A., 2014. The Effects of Conditional Cash Transfers on Social Engagement and Trust in Institutions: Evidence from Peru's Juntos Program (October 31). Discussion Paper 24/2014, German Development Institute, Bonn, Available at SSRN: http://ssrn.com/abstract=2512666 or http://dx.doi.org/10.2139/ssrn.2512666

Camacho, A., Mejía, D., 2013. Las externalidades de los programas de transferencias condicionadas sobre el crimen: El caso de Familias en Acción en Bogotá. Inter-American Development Bank. Working paper IDB-WP-406.

Camerer, C., Fehr, E., 2004. Measuring social norms and preferences using experimental games: A guide for social scientists. In: Henrich, J., Bowles, S., Boyd, R., Camerer, C., Fehr, E., Gintis, H. (Eds.), Foundations of Human Sociality: Economic Experiments and Ethnographic Evidence from Fifteen Small-Scale Societies. Oxford University Press, Oxford.

Cameron, L., Shah, M., 2014. Can mistargeting destroy social capital and stimulate crime? Evidence from a cash transfer program in Indonesia. Econ. Dev. Cultural Change 62 (January (2)), 381–415.

Candelo, N., Polania-Reyes, S., 2008. Pasos metodológicos de un diseño experimental para medir capital social y acción colectiva en seis ciudades latinoamericanas (Methodological steps for an experimental design to measure social capital and collective action in six Latinamerican cities). Working paper CEDE (17) Universidad de los Andes.

Capra, C.M., Lanier, K., Meer, S., 2008. Attitudinal and behavioral measures of trust: a new comparison. Working paper Emory University.

Cárdenas, J.C., 2003. Real wealth and experimental cooperation. J. Dev. Econ. 70, 263–289.

Cárdenas, J.C., Chong, A., Ñopo, H., Horowitz, A.W., Lederman, D., 2009. To what extent do Latin Americans Trust, Reciprocate, and Cooperate? Evidence from experiments in six Latin American Countries [with comments]. Economia (2), 45–94.

Cárdenas, J.C., Chong, A., Ñopo, H., 2013. Stated social behavior and revealed actions: evidence from six Latin American countries. J. Dev. Econ. 104, 16–33.

Cárdenas, J.C., Jaramillo, C., 2007. Cooperation in large networks: an experimental. In: Documentos CEDE. Universidad de los Andes-Cede, Bogota.
Carpenter, J., 2002. Measuring social capital: adding field experimental methods to the analytical toolbox. In: Ramaswamy, S., Isham, J., Kelly, T. (Eds.), Social Capital, Economic Development and the Environment. Edward Elgar.
Carpenter, J., Daniere, A., Takahashi, L., 2004. Cooperation, trust and social capital in Southeast Asian urban slums. J. Econ. Behav. Organ. 55 (4), 533–552.
Casey, K., Glennerster, R., Miguel, M., 2012. Reshaping institutions: evidence on aid impacts using a preanalysis plan. Quart. J. Econ. 127 (4), 1755–1812, http://dx.doi.org/10.1093/qje/qje027.
Chattopadhyay, R., Duflo, E., 2004. Women as policy makers: evidence from a randomized policy experiment in India. Econometrica 72 (5), 1409–1443.
Christoforou, A., Davis, J.B. (Eds.), 2014. Social Capital and Economics. Routledge, London and New York.
Coleman, J., 1988. Social capital in the creation of human capital. Am. J. Sociol. 94, S95–S120.
Coleman, J.S., 1990. Foundations of Social Theory. Harvard University Press, Cambridge, MA and London, England.
Cohen, E., Franco, R. (Eds.), 2006. Transferencias con Corresponsabilidad. Una Mirada Latinoamericana. México. Fac. Ciencias Soc. (FLACSO), p. 434.
Croson, R.T.A., 1996. Contributions to Public Goods: Altruism or Reciprocity? The Wharton School, University of Pennsylvania.
De La O, A.L., 2013. Do conditional cash transfers affect electoral behavior? Evidence from a randomized experiment in Mexico. Am. J. Polit. Sci. 57 (1), 1–14.
DNP, 2010, July. The journey: ten years of Familias en Acción (El camino recorrido: Diez Años de Familias en Acción). Acción Social Departamento Nacional de Planeación, Bogota, Colombia.
ECLAC United Nations Economic Commission for Latin America and the Caribbean, 2005. In: Arriagada, I. (Ed.), Aprender de la experiencia: el capital social en la superación de la pobreza, vol. 86. United Nations Publications.
Fearon, J., Humphreys, M., Weinstein, J.M., 2009. Can development aid contribute to social cohesion after civil war? Evidence from a field experiment in post-conflict liberia. Am. Econ. Rev. 99 (2).
Fiszbein, A., Schady, N.R., Ferreira, F.H.G., Grosh, M., Kelleher, N., Olinto, P., Skoufias, E., 2009. Conditional Cash Transfers: Reducing Present and Future Poverty. World Bank Publications.
Gächter, S., Hermann, B., Thoni, C., 2004. Trust, voluntary cooperation, and socio-economic background: survey and experimental evidence. J. Econ. Behav. Organ. Elsevier 55 (December (4)), 505–531.
Gilligan, M.J., Pasquale, B.J., Samii, C., 2013. Civil war and social cohesion: lab-in-the-field evidence from Nepal. Am. J. Polit. Sci.
Grossman, S.J., Hart, O., 1980. Takeover bids, the free-rider problem, and the theory of the corporation. Bell J. Econ. 11 (1), 42–64 (Spring, 1980).
Jones, B.F., Olken, B.A., 2005. Do leaders matter? National leadership and growth since World War II. Quart. J. Econ., 835–864.
Karlan, D., 2005. Using experimental economics to measure social capital and predict financial decisions. Am. Econ. Rev. 95 (5), 1688–1699.
Labonne, J., 2013. The local electoral impacts of conditional cash transfers: evidence from a field experiment. J. Dev. Econ. 104 (September), 73–88.
Latham, G.P., Saari, L.M., 1979. Importance of supportive relationships in goal setting. J. Appl. Psychol. 64 (2), 151.
Latorre López, M.C., 2004. Sobre la relación positiva entre el capital social y la violencia urbana: un análisis teórico y empírico (About the positive relationship between social capital and urban violence: a theoretical and empirical analysis]). In: Documentos CEDE no. 36. Bogota, DC.
Ledyard, J.O., 1995. Public goods: a survey of experimental research. In: Roth, A.E., Kagel, J. (Eds.), The Handbook of Experimental Economics. Princeton University Press, Princeton, NJ, pp. 111–194.
MacAuslan, I., Riemenschneider, N., 2011. Richer but resented: what do cash transfers do to social relations? IDS Bull. 42 (6), 60–66, http://onlinelibrary.wiley.com/doi/10.1111/j.1759-5436.2011.00274.x/pdf
Macours, K., Vakis, R., 2014. Changing households' investment behaviour through social interactions with local leaders: evidence from a randomised transfer program. Econ. J. 124, 607–633, http://dx.doi.org/10.1111/ecoj.12145.
Manacorda, M., Miguel, E., Vigorito, A., 2011. Government transfers and political support. Am. Econ. J. Appl. Econ. 3 (3), 1–28.
Mansuri, G., Rao, V., 2012. Localizing Development: Does Participation Work? World Bank Publications.
MESEP, 2012. Mission for Splicing Series Employment, Poverty and Inequality (MESEP). Monetary poverty in Colombia: New methodology and figures 2002–2010 (Misión para el Empalme de las Series de Empleo, Pobreza y Desigualdad (mesep). Pobreza monetaria en Colombia: Nueva metodología y cifras 2002–2010). In: Report. DNP, DANE. Bogota, Colombia.
Nguyen, T.C., Rieger, M., 2014. Community-driven development and social capital: Evidence from Morocco. EUI Working Paper MWP 2014/02.
Nupia, O., 2012. Anti-poverty programs and presidential election outcomes: Familias en acción in colombia. Universidad de los Andes (2012). Revised version of Documento CEDE 14 (2011).
Ñopo, H., Cárdenas, J.C., Polania-Reyes, S., Candelo, N., Calónico, S., Chong, A., 2008. To What Extent Do Latin Americans Trust and Cooperate? Field Experiments on Social Exclusion in Six Latin American Countries. In: IADB IPES 2008 Report: Outsiders? The Changing Patterns of Exclusion in Latin America and the Caribbean (David Rockefeller/Inter-American Development Bank) (Chapter 7).
Olson, 1965. The Logic of Collective Action. Harvard University Press, Cambridge.
Ostrom, E., 1998. A behavioral approach to the rational choice theory of collective action. Am. Polit. Sc. Rev. 92 (1), 1–22.
Pellerano, L., 2008. Assessing the impact of conditional cash transfer programs on social participation: evidence from Colombia', Institute for Fiscal Studies, Manuscript.
Pellerano, L., Moratti, M., Jakobsen, M., Bajgar, M., Barca, V., 2014. Child Grants Program Impact Evaluation. Follow-up Report. April, Available from www.fao.org/fileadmin/user_upload/p2p/Documents/draft_CGP_Follow_Up_v11_out.pdf
Polania-Reyes, S., 2005. Capital Social E Ingreso De Los Hogares Del Sector Urbano En Colombia (Social Capital and Household Income in Urban Colombia). Rev. Desarro. Soc. Univ. Andes 50 (2).
Pop-Eleches, C., Pop-Eleches, G., 2012. Targeted government spending and political preferences. Quart. J. Polit. Sci. 7 (3), 285–320.
Putnam, R.D., 1993. Making Democracy Work: Civic Traditions in Modern Italy. Princeton University Press, Princeton, NJ.
Putnam, R.D., 1995. Bowling alone: America's declining social capital. J. Democr. 6, 68.
Rabin, M., 1993. Incorporating fairness into game theory and economics. Am. Econ. Rev. 83 (5), 1281–1302.
Samuelson, P.A., 1954. The Pure Theory of Public Expenditure. Rev. Econ. Stat. 36 (4), 387–389.
Tajfel, H., 1982. Social psychology of intergroup relations. Annu. Rev. Psychol. 33 (1), 1–39.
Uphoff, N., 1999. Understanding social capital: learning from the analysis and experience of participation. In: Dasgupta, P., Serageldin, I. (Eds.), Social Capital. A Multifaceted Perspective. Banco Mundial, Washington, DC.
Voors, M.J., Nillesen, E.E.M., Verwimp, P., Bulte, E.H., Lensink, R., Van Soest, D.P., 2012. Violent conflict and behavior: a field experiment in Burundi. Am. Econ. Rev. 102 (2), 941–964.
Warren, M.R., 1998. Community building and political power a community organizing approach to democratic renewal. Am. Behav. Sci. 42 (1), 78–92.
Wong, S., 2012. What have been the impacts of World Bank Community-Driven Development Programs? CDD impact evaluation review and operational and research implications. World Bank, Washington, DC.
Zárate, R.A., Conover, E., Camacho, A., Báez, J.E., 2013. Conditional Cash Transfers, Political Participation, and Voting Behavior, Available from: http://cie.itam.mx/englishversion/faculty/fea_elect_Adriana%20Camacho.pdf
Zucco, C., 2011. Cash-Transfers and Voting Behavior: An assessment of the political impacts of the Bolsa Familia Progam, mimeo. Princeton University.

Economic Incentives and Social Preferences: Substitutes or Complements?

Samuel Bowles and Sandra Polanía-Reyes*

Explicit economic incentives designed to increase contributions to public goods and to promote other pro-social behavior sometimes are counterproductive or less effective than would be predicted among entirely self-interested individuals. This may occur when incentives adversely affect individuals' altruism, ethical norms, intrinsic motives to serve the public, and other social preferences. The opposite also occurs—crowding in—though it appears less commonly. In the fifty experiments that we survey, these effects are common, so that incentives and social preferences may be either substitutes (crowding out) or complements (crowding in). We provide evidence for four mechanisms that may account for these incentive effects on preferences: namely that incentives may (i) provide information about the person who implemented the incentive, (ii) frame the decision situation so as to suggest appropriate behavior, (iii) compromise a control averse individual's sense of autonomy, and (iv) affect the process by which people learn new preferences. An implication is that the evaluation of public policy must be restricted to allocations that are supportable as Nash equilibria when account is taken of these crowding effects. We show that well designed fines, subsidies, and the like minimize crowding out and may even do the opposite, making incentives and social preferences complements rather than substitutes. (JEL D02, D03, D04, D83, E61, H41, Z13)

1. Introduction

Thomas Schelling recalls his "exciting and stimulating times" in the early 1950s as a young staffer in the Executive Office of the President. "People worked long hours," he remembered in a recent communication to one of us, "and felt compensated by the sense of accomplishment and . . . personal importance. Regularly a Friday afternoon

meeting at the White House would go on until 8 or 9, when the chairman would suggest resuming Saturday morning. Nobody demurred. We all knew it was important, and we were important. . . . What happened when the President issued an order that anyone who worked on Saturday was to receive overtime pay…? Saturday meetings virtually disappeared."

Was Schelling's experience atypical? Incentives work, often affecting the targeted behavior almost exactly as conventional economic theory predicts: textbook examples include the work response of Tunisian sharecroppers and American windshield installers, and the reduced criminal activities of those former Italian convicts who could expect more severe sentences if convicted again (Laffont and Matoussi 1995; Lazear 2000; Drago, Galbiati, and Vertova 2009).

But explicit economic incentives sometimes have surprisingly limited effects, and may even be counterproductive. Substantial rewards for high school matriculation in a randomized experiment in Israel had no impact on boys and little effect on girls except among those already quite likely to matriculate (Angrist and Lavy 2009). Large and, in most cases, immediate cash payment in return for tested scholastic achievement in 250 urban schools in the United States were almost entirely ineffective, while incentives for student inputs (reading a book, for example) had the intended, if modest effects (Fryer 2011). In an unusual natural experiment, the imposition of fines designed to reduce hospital stays in Norway had the opposite effect (Holmås et al. 2010), while in England hospital stays were greatly reduced by a policy designed to evoke shame and pride in hospital managers rather than the calculus of profit and loss (Besley, Bevan, and Burchardi 2009).

Anecdotal accounts of what appear to be even more dramatic cases of counterproductive incentives are common. On December 1, 2001, the Boston Fire Department terminated its policy of unlimited paid sick days, replacing it with a fifteen-day sick day limit; pay would be docked for firemen exceeding the limit. The firemen responded to the new incentives: those calling in sick on Christmas and New Year's Day increased tenfold over the previous year. The Fire Commissioner retaliated by canceling their holiday bonus checks (Belkin 2002). The firemen were unimpressed: the year following they claimed 13,431 sick days; up from 6,432 the previous year (Greenberger 2003). Many of the firemen, apparently angered by the new system, abused it or abandoned their previous ethic of serving the public even when injured or not feeling well.

Not surprisingly, then, since Richard Titmuss's *The Gift Relationship: From Blood Donations to Social Policy*, economists have been intrigued by the claim that policies based on explicit economic incentives may be counterproductive when they induce people to adopt what Titmuss called a "market mentality" or in some other way compromise preexisting values to act in socially beneficial ways. But few were persuaded (Solow 1971; Arrow 1972; Bliss 1972).

At the time of its publication, there were two strong reasons to doubt Titmuss's claim: there was little hard evidence that social preferences are important influences on individual behavior; and there was even less evidence (in the Titmuss 1971 book or elsewhere) that social preferences would be undermined by explicit economic incentives (which we will call simply "incentives" without the adjectives, meaning interventions to influence behavior by altering the economic costs or benefits of some targeted activity).

Theoretical and empirical advances over the intervening years provide the basis for a reconsideration of these issues (Kreps 1997;

Elster 1998; Rabin 1998; Loewenstein 2000; Sobel 2002; surveyed in Meier 2007b). First, evidence from both the behavioral experimental laboratory and the field is consistent with the view that social preferences are important influences on economic behavior (Fehr, Gächter, and Kirchsteiger 1997; Bewley 1999; Fehr and Schmidt 1999; Fehr and Gächter 2000; Young and Burke 2001; Bandiera, Barankay, and Rasul 2005; Falk, Fehr, and Fischbacher 2005; Fehr, Klein, and Schmidt 2007; DellaVigna 2009; Leider et al. 2009; Sloof and Sonnemans 2011).

Second, the importance of incomplete contracts has been widely recognized in theoretical works and studied empirically (Stiglitz 1987; Laffont and Matoussi 1995; Tirole 1999). Partly as a result, the terms trust, reciprocity, fairness, gift exchange, and social capital now appear in the modeling and empirical study of principal–agent relationships, the provision of public goods, and other standard economic applications, often referring to the social norms and identities that underwrite mutually beneficial exchange in the absence of complete contracts (Arrow 1971; Becker 1976; Akerlof 1984; Helsley and Strange 2000; Benabou and Tirole 2006; MacLeod 2007; Sliwka 2007; Ellingsen and Johannesson 2008; Akerlof and Kranton 2010).

Third, advances in the theory of public policy have addressed cases in which incentives affect both beliefs and preferences and may thus have unintended effects (Lucas 1976; Taylor 1987; Bowles 1989; Aaron 1994; Sunstein 1996; Frey 1997; Cooter 2000; Bowles 2004; Bar-Gill and Fershtman 2005; Sobel 2005; Cervellati, Esteban, and Kranich 2010).

Finally, a substantial literature (surveyed in Frey and Jegen 2001, Bowles 2008, and Gneezy, Meier, and Rey-Biel 2011) emerged showing that explicit incentives are sometimes counterproductive.

2. *Overview: Incentives and Social Preferences as Substitutes or Complements*

We use the term "social preferences" to refer to motives such as altruism, reciprocity, intrinsic pleasure in helping others, inequity aversion, ethical commitments, and other motives that induce people to help others more than would an own-material-payoff maximizing individual. Our use of the term is thus not restricted to cases in which the actor assigns some value to the payoffs received by another person, as in the utility functions of Fehr and Schmidt (1999), Rabin (1993), and Levine (1998). While these functions provide a convenient way to model some of the motivations for pro-social behavior, we use the broader definition because moral, intrinsic, or other reasons unrelated to a concern for another's payoffs often motivate people to help others, adhere to social norms, and act in other pro-social ways even when it is personally costly to do so. A person, for example, may adhere to a social norm not because of the harm that a transgression would do to others, but because of the kind of person she would like to be; helping the people who are homeless may be motivated by Andreoni's "warm glow" of giving rather than a concern with the wellbeing of the poor (Andreoni 1990).

The standard (if generally implicit) assumption in economics is that the behavioral functions relevant for mechanism design, public economics, and related fields are separable in social preferences (should they exist) and incentives. This means, for example, that the citizen's response to variations in a subsidy for contributions to a public good is independent of her preexisting level of social preferences. It also means that the effect of variations in her preexisting noneconomic motivations on the citizen's level of contributions does not depend on the presence or magnitude of incentives.

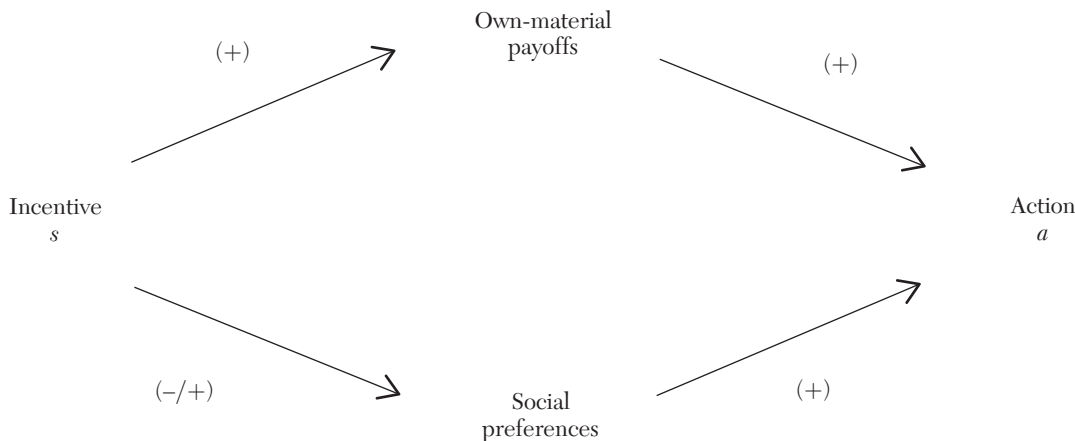We call this the separability assumption. (A recent example is Levitt and List 2007.) It is

Own-material
payoffs

(+)

(+)

Incentive
s

Action
a

(−/+)

Social
preferences

(+)

Figure 1. Crowding Effects of Incentives

Notes: The direct and indirect effects of incentives on contribution to a public good (*a*). The effect of an incentive (*s*) on social preferences may be either to alter their behavioral salience for the action (social preferences are state-dependent) or to affect the manner in which preferences are updated, thereby altering the individual's social preferences (endogenous social preferences). Crowding out occurs when the effect of an incentive on social preferences is negative (assuming that the effect of social preferences on the action is positive, as shown). Crowding in (the opposite) also occurs.

a useful simplification for problems in which social preferences can safely be ignored, as in the choice of apples versus oranges and other standard textbook examples. But it can also be misleading as it implies that taxes, subsidies, and other incentives affect behavior only by altering the economic costs and benefits of the targeted activities. We will present evidence that the separability assumption often does not hold, and as a result social preferences may be either heightened by incentives appealing to self-interest or, the more commonly observed case, affected adversely by incentives.

This is illustrated in figure 1 where, due to the effect of incentives on preferences, the total—direct and indirect—effect of the incentive may fall short of that which works directly on the costs and benefits of the targeted activity. In this case, we say that incentives crowd out social preferences and that incentives and social preferences are

substitutes: the effect of each on the targeted activity declines, the greater is the level of the other. Where the effect on social preferences is positive, crowding in occurs and social preferences and incentives are complements, the level of each enhancing the effect of the other.

The possibility that incentives designed for material payoff-maximizers might have adverse effects is a familiar theme in political science (Taylor 1987; Grant 2011), psychology (Deci 1975), sociology (Healy 2006), and the other social sciences; but it has found few adherents in economics. The reason is that we have adopted a simplifying strategy that goes back at least to John Stuart Mill (1867[1848], 97): "[Political economy] does not treat of the whole of man's nature . . . it is concerned with him solely as a being who desires to possess wealth, . . . it predicts only such . . . phenomena . . . as take place in consequence of the pursuit of wealth. It makes

entire abstraction of every other human passion or motive." In other words, we ignore the two lower arrows in figure 1. But recent experimental and other evidence has prompted many economists to reconsider Mill's simplification. To further this reconsideration we here provide a taxonomy and two models of incentive effects on preferences based on two distinctions: their nature and their causes.

Concerning the first, people often react to the mere presence of incentives rather than their extent: giving to charity when tax breaks are involved (whatever their magnitude) may feel different or send a different signal than would be the case in the absence of these incentives. But the extent of the incentive may also matter (Gneezy 2003). Thus the effects of incentives on social preferences may be either marginal (depending continuously on the level of the incentive) or categorical (the presence of incentives affecting social preferences independently of their level) or a combination of the two.

We also distinguish between two causes of incentive effects on preferences. First, behavior is acutely sensitive to the nature of the decision situation (Ross and Nisbett 1991; Tversky and Kahneman 1981); and, as we will see, the presence or extent of incentives provides information about the situation. A psychologist might say that preferences are "situation-dependent" and that incentives provide situational clues. We say that the preferences are state-dependent, with differing incentives constituting different states. In the next section we offer a model of incentive-state-dependent preferences and provide data indicating that both categorical and marginal crowding out occurs.

State dependence arises because actions are motivated by a heterogeneous repertoire of preferences—from spiteful to payoff-maximizing to generous, for example—the salience of which depends on the nature of the

decision situation—interacting with a domineering supervisor, shopping, or relating to one's neighbors, for example. To see how this works, think about gifts. Economists know that money is the perfect gift—it replaces the giver's less well-informed choice of a present by the recipient's own choice. But when the holidays come around few economists give money to their friends, family and colleagues. This is because we also know that money cannot convey thoughtfulness, concern, whimsy, or any of the other messages that nonmonetary gifts sometimes express. A gift, we know, is more than a transfer of resources; it is a signal about the giver and her relationship to the recipient, and money changes the signal.

Can the same be said of incentives? A long tradition in psychology has concluded that it can: "The multiple meanings of . . . tangible rewards are reflected in our everyday distinction among bribes and bonuses, incentives and salaries. . . . they carry different connotations concerning, for example, (*i*) the likely conditions under which the reward was offered, (*ii*) the presumed motives of the person administering the reward, and (*iii*) the relationship between the agent and the recipient of the reward" (Lepper et al. 1982, italic numbers added).

As Lepper and his coauthors say, incentives may affect preferences for a reason familiar to economists, that is because they indicate "the presumed motives of the person administering the reward." By implementing an incentive, a principal reveals information about his or her intentions (own payoff maximizing versus fair-minded, for example) as well as beliefs about the target of the incentives (hardworking or not) and the targeted behavior (how onerous it is, for example.) This information, in turn, may then affect the target's noneconomic motivation to undertake the task at hand. In section 5, we present experimental evidence that the information provided about the principal

can sometimes attenuate or even reverse the intended effect of the incentive. Of course when an incentive provides good news about the principal's intentions or type—when rewards are offered, for example, rather than fines—it may recruit the target's social preferences to work synergistically with the direct effect of the incentive. In this case, incentives and social preferences become complements rather than substitutes. We will see (in section 9 and 11) that this crowding in phenomenon is sometimes observed in experiments, for example, when the principals implementing incentives are peers in a Public Goods Game who pay to fine free riders in order to support cooperative norms.

But there are other reasons, less familiar to economists, for state-dependence: reasons that do not concern information about the principal, and that may be at work even in nonstrategic settings. A second mechanism is that incentives provide cues about (as Lepper and his coauthors put it) "the likely conditions under which the reward was offered": by framing a decision situation, economic incentives may provide cues for appropriate behavior. This second mechanism is distinguished from the first in the experimental evidence by the fact that in the former the incentives are implemented by a principal who is a player in the game; while in the latter the targets of the incentive are not playing against the incentive designer; rather the incentives are introduced by the experimenter.

Situational cues may be very subtle, and our responses to them unwitting. When experimental subjects had the opportunity to cheat on a test and as a result to gain higher monetary rewards, less than quarter did so when the room was brightly lit, but more than half cheated when the room was slightly less well lit (the variations in lighting had no effect on the observability of cheating). In another experiment, subjects

who wore (nonprescription) dark glasses were much less generous to their partner in a Dictator Game than were those outfitted with clear glasses (Zhong, Bohns, and Gino 2010). The dark glasses and darkened room gave the subjects a sense of anonymity, the researchers found. But it was entirely illusory: it is difficult to imagine that a subject could really think that his own wearing dark glasses would make him less observable, especially given that the experiment was conducted at computer terminals in closed cubicles.

The degree of anonymity differs dramatically as we move between family, workplace, marketplace and other domains of social interaction. Fiske (1992) provides a taxonomy of four psychological models corresponding to distinct kinds of social relationships: authoritarian, communal, egalitarian and market, each with culturally prescribed patterns of appropriate behavior. Depending on the information they convey, incentives may signal that the situation corresponds to any one of these four types, and therefore evoke distinctive responses.

We will see that a plausible explanation of some of the framing effects of incentives observed in experiments is that it occurs because market-like incentives trigger what psychologists term "moral disengagement" (Bandura 1991), a process that occurs because "people can switch their ethicality on and off" (Shu, Gino, and Bazerman 2011, 31). In section 6, we review experiments in which crowding out appears to have been the result of moral disengagement. Depending on the information they convey, incentives may also trigger the opposite—moral engagement—and, as we will see in section 9, experiments provide a few examples of this form of crowding in, illustrating the possible synergy or complementarity between social preferences and incentives.

The third mechanism that makes social preferences state-dependent is the crowding

out of intrinsic motives by incentives (or constraints) that compromise a subject's sense of autonomy (Deci and Ryan 1985; Deci, Koestner, and Ryan 1999). These effects may occur in strategic situations where the bad news that incentives convey concerns the desire of a principal to control the agent. But most of the experimental evidence for this third crowding out mechanism comes from nonstrategic settings (the experimenters, not a principal implements the incentive). The underlying psychological mechanism appears to be a fundamental desire for "feelings of competence and self-determination" that are associated with intrinsically motivated behavior (Deci 1975).

According to this interpretation, where people derive pleasure from an action per se in the absence of other rewards, the introduction of incentives may "overjustify" the activity and reduce the individual's sense of autonomy. This self-determination mechanism differs from the previous two mechanisms—bad news about a principal and moral disengagement—because it arises from the target's desire for autonomy and does not depend on the target inferring negative information about a principal or clues about appropriate behavior. This is particularly evident in some early "overjustification" experiments in which when a financial reward was offered by the experimenter, children often forsook previously uncompensated activities in which they had enthusiastically engaged, like painting. More recent experiments show the same negative effects of incentives on altruistic behavior (Warneken and Tomasello 2008): in the absence of rewards, children less than two years old avidly helped an adult retrieve an out of reach object; but after being rewarded with a toy for their helping behavior the helping rate fell by forty percent.

The fact that the incentive was a reward rather than a penalty suggests that it did not convey negative information about the incentive designer, but instead altered the meaning of the activity itself from one that expressed autonomy to one that expressed compliance. The interpretation that self-determination is involved in the negative response to incentives is consistent with the fact that close supervision or arbitrary temporal deadlines for completion of an otherwise enjoyable activity have effects very similar to financial rewards (Lepper et al. 1982). In section 7, we survey experimental evidence for this "control aversion" mechanism for state-dependent preferences.

We have just described three (partially overlapping) reasons why the state-dependent nature of preferences might lead to crowding out. For ease of reference, we will call them "bad news," "moral disengagement," and "control aversion." But, in addition to incentives altering the preferences that motivate an individual's action by altering the subject's sense of the situation, there is a second and quite different way that incentives may affect preferences. The type and extent of a society's use of economic incentives also may affect the process of preference-updating by which individuals acquire new tastes or social norms that will persist over long periods. Models from biology, anthropology, and economics allow us to formalize this learning process (Cavalli-Sforza and Feldman 1981; Boyd and Richerson 1985; Güth and Kliemt 1994; Bowles 1998; Bisin and Verdier 2011; Bowles and Gintis 2011).

The key difference between endogenous and state-dependent preferences is that, in the former case, the effect of the incentive on preferences persists in the long run because the updating process on which cultural transmission is based typically occurs during youth and its effect endures over decades if not entire lifetimes. We say that incentives affect preferences in both the state-dependent and endogenous preference case, but the mechanism of the effect is different: in the former case the incentive is a

reversible signal about the principal or the situation, in the latter the incentive alters the preference-updating process.

An example unrelated to incentives may clarify the difference between endogenous and state-dependent preferences. As Italian residents, your authors now eat a lot more pasta than we did in our countries of origin. Abstracting from possible international price differences, this could be another case of "when in Rome, do as the Romans." Or it might be that we have newly come to enjoy the taste of pasta, perhaps through extensive exposure to it while in Italy. Which case it is—state-dependent or endogenous preferences—would be revealed by what we will eat back in Bogotá or Santa Fe. If we go back to *arepas* or potatoes, then our taste for pasta was state-dependent. If we remain *pastaphiles*, then our preferences have endogenously changed.

Preferences may be endogenous in this sense because the extent to which a society relies on economic incentives—as opposed to other kinds of motivations and controls—may affect how people learn new preferences (evidence for the endogeneity of preferences is surveyed in Bowles (1998, 2004) and Bowles and Gintis (2011). The learning on which preference endogeneity is based is of course a long-term process unlikely to be observed in a brief experiment. Nonetheless, experiments may provide clues that learning is affected: we take as evidence consistent with preference endogeneity those cases in which crowding effects of incentives persist after the removal of the incentive. (Other explanations not involving endogenous preferences are generally also possible in these situations.)

In sections 3 and 4, we make explicit the underlying causal mechanisms through the use of models of state-dependent and endogenous preference formation. Table 1 summarizes the differences.

Our empirical strategy (based on experimental results) is to observe the total effect of incentives on behavior and to note whether this differs from the predicted direct effect (the top arrows in figure 1) in order to infer the effects of incentives on (unobserved) social preferences and thereby on actions (the bottom two arrows). Our data set includes all the economic experiments we have been able to locate that allow a test of the separability assumption. Our tables include more than a hundred different subject pools, over twenty-six thousand subjects from thirty-six countries, playing Dictator, Trust, Ultimatum, Public Goods, Third Party Punishment, Common Pool Resource, Gift Exchange, and other principal–agent games. These are all settings in which one's actions affect the payoffs of others so that social preferences may affect a subject's experimental behavior. We find evidence of nonseparability in all of these games. Because nonseparability, as we will see, arises from the social relationships among those imposing incentives and their targets and the nature of the incentive, and because game structures differ in this respect, it would be surprising if the nature and degree of non separability did not differ across these games. However, lacking a metric for nonseparability that is comparable across games, we have not explored this possibility.

Few experiments have thus far been designed to address the causes of nonseparability, so the inferences that we draw must be provisional. The experimental methods that have become standard in economics include playing for real stakes, excluding deception, and making explicit use of game theoretic concepts to clarify the role of incentives. As experimental methods differ considerably across disciplines, and for reasons of space, we limit the entries in the tables to experiments done by economists. We refer to a number of important experiments done using other methods in the text. All reported

TABLE 1
ECONOMIC INCENTIVES AND SOCIAL PREFERENCES:
ENDOGENOUS AND STATE-DEPENDENT EFFECTS AND MECHANISMS

Source and Characterization (modeled in section §)	Mechanisms	Description (§: section with empirical evidence relevant to this mechanism)
State-dependent preferences Incentives affect the *behavioral salience of an individual's social preferences*. §3	Information "bad news"	Incentive signals the designer's type or beliefs about the target or the nature of the targeted task, and may convey illegitimate pursuit of self-interest by principal. §5
	Framing "moral disengagement"	Incentive signals the type of situation and hence appropriate behavior for the target, and may activate own payoff-maximizing modes of thought. §6
	Self-determination "control aversion"	Incentive affects target's sense of autonomy, and may signal unacceptable control and motivate resistance. §7
Endogenous preferences Incentives affect the *environment in which preferences are learned* and therefore the *stationary distribution of preference types in the population* (i.e. the fraction of population with social preferences). §4	Conformist preference-updating	Incentives reduce the perceived population fraction of social preference types. The extent to which a society relies on economic incentives—as opposed to other kinds of motivations and controls—will affect how people learn new preferences that may persist over long periods. §8

Notes: As a result of the mechanisms listed incentives and social preferences may be either complements (crowding in) or substitutes (crowding out). In the conclusion we consider cases in which the degree of endogenous or state-dependent nonseparability is subject to public policy because the crowding parameters introduced below (λ_m, λ_c, Λ_m, and/or Λ_c) may themselves be affected by incentives. Additional mechanisms for endogenous crowding out are provided in Bowles (2004).

results are statistically significant at conventional levels unless noted.

Incentives may have counterintuitive and counterproductive effects for reasons other than nonseparability (Seabright 2009). Strong monetary incentives, for example, may overmotivate an agent leading to greater than the optimal level of arousal. This appears to be the mechanism underlying the negative effects of high incentives found in three experiments by Ariely et al. (2009). In other cases, incentives alter individuals' beliefs about the actions of others, with possibly counterintended effects. This is often thought to be the case when tax authorities announce stiff penalties for underpayment, unwittingly letting the public know that cheating is common and thereby promoting rather than deterring it. We do not consider these and other cases of counterproductive incentives where the mechanisms are unrelated to the nonseparability of incentives and social preferences, which is the focus of this paper.

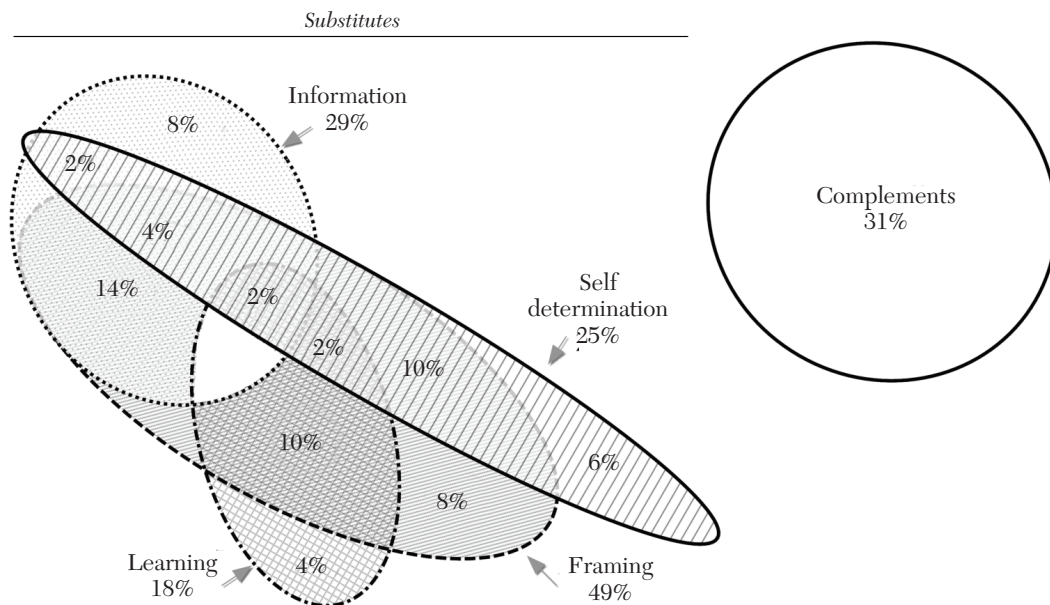Some of the experimental results presented below may be explained by more than

Figure 2. Summary of Experimental Evidence on the Four Crowding Out Mechanisms and Crowding In

Notes: In the figure on the left, the mechanisms accounting for crowding out are shown. The intersections show cases in which more than one mechanism may be involved. For example 14 percent of the experiments are consistent with both the framing and information about the incentive designer mechanisms. The circle in the upper right refers to crowding in (we have not separated out the mechanisms in this case.) The numbers indicate the percentage of the total of fifty studies that exhibit the mechanisms indicated. There are no studies in the intersections that are blank.

one of our four mechanisms that account for non-separability, either because the mechanisms are not mutually exclusive so that multiple mechanisms are at work, or because the experiment does not provide sufficient information to say which mechanism accounts for the evidence of nonseparability. In these ambiguous cases, we classified the experiment as an illustration of the mechanism which we thought best accounted for the crowding result that we report (We indicate in each table where such ambiguities occur and which other mechanisms may have been at work.) As a preview, figure 2 presents a summary of our findings, the size of the ellipses indicating the percentage of studies

that exhibit each of the four crowding out mechanisms in question, and the intersections giving the cases where multiple mechanisms may be involved.

There are two reasons why, despite the considerable number of experiments in which incentives appear to affect social preferences, it is difficult to estimate how prevalent these effects are in real economies. First, the experimental games involved are about social dilemmas or sharing with others, that is, settings in which social preferences are likely to be important and therefore there is something to be crowded out or in. While the experimental evidence suggests that crowding out may affect blood donations or

participation in community service projects, it does not have much to say about the effect of incentives on shopping behavior or cleaning hotel rooms. Second, while section 10 presents evidence that experimental play in these social dilemmas predicts behavior in some nonexperimental situations, isolating social preferences from other influences on behavior in natural settings is difficult. We conclude in sections 11 and 12 with policy implications.

3. *Incentives as Signals: A Model of State-Dependent Preferences*

In this section (following Hwang and Bowles 2011a and Bowles and Hwang 2008), we model incentive effects on state-dependent preferences and clarify the distinction between categorical and marginal incentive effects by means of an empirical illustration. We consider an individual who may bear a cost to take an action that confers benefits on others, which may be encouraged by a subsidy implemented by a social planner. Citizens also have values that may motivate such pro-social actions even in the absence of the subsidy. We study a single member of a community of identical citizens who may contribute to a public project by taking an action a at a cost $g(a)$ that is increasing and convex in its argument, and that may be offset partially by a subsidy s, that is proportional to the individual's level of contribution. The output of the project is available in equal measure to all, and it varies positively and linearly with A, the sum of the n members' contributions, according to φA where φ is a positive constant.

We express the individual's social preferences as v, the effect of an increase in the contribution level on the individual's utility that is unrelated to material payoffs. Thus we have the individual's utility

$$(1) \quad u = \varphi A - g(a) + as + av.$$

We make explicit the sources of nonseparability by the value function:

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

where the indicator $\mathbf{1}\{s > 0\} = 1$ if $s > 0$ and zero otherwise. In equation (2), $\lambda_0 \geq 0$ measures the citizen's baseline social preferences namely the citizens values in the absence of a subsidy, λ_c (which may be of either sign) measures the categorical effect of the presence of an incentive, and λ_m (which also may be of either sign) measures the marginal effect of variations in s on values for $s > 0$. The crowding effects represented by λ_c and λ_m in (2) may arise because of any of the three mechanisms by which state-dependent preferences arise: bad news, moral disengagement, or control aversion.

The individual's utility is thus

$$(3) \ u = \varphi A - g(a)$$

$$+ \ a(s + \lambda_0(1 + \mathbf{1}\{s > 0\}\lambda_c + s\lambda_m))$$

and the individual's utility maximizing contribution (a^*) equates the marginal cost of contributing to the marginal benefits (the returns from the public good plus the subsidy plus the effect on the individual's values), or:

$$(4) \quad g'(a^*) = \varphi + s$$

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

We assume that in the absence of a subsidy the contributions of the citizens to the public good given by (4) are inefficient in the sense that there exists a mutual increase in contributions that would make all citizens better off. The causal structure of the model is illustrated in figure 3.

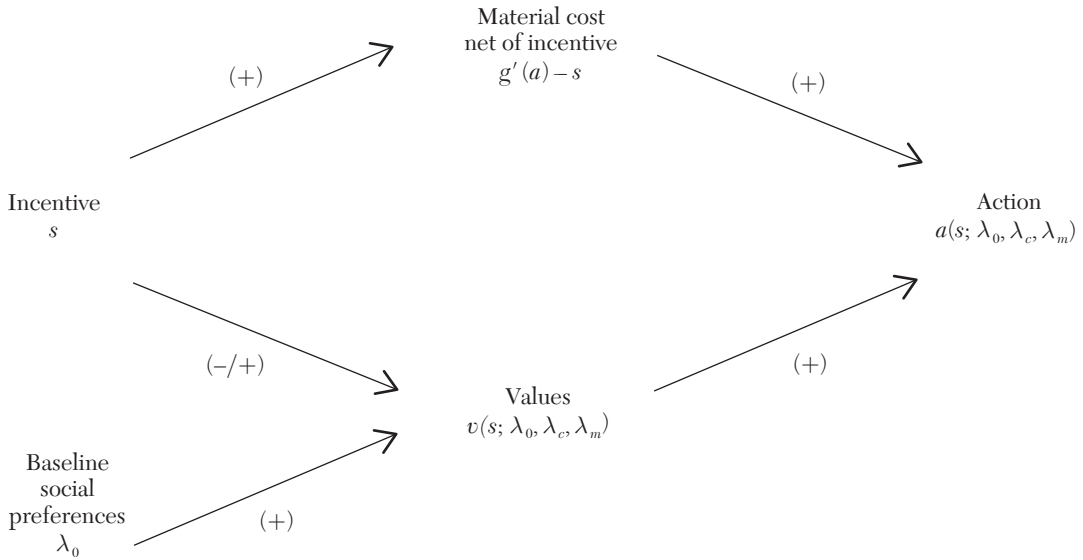The introduction of a subsidy increases contributions by raising the marginal

Figure 3. Crowding Effects of Incentives for an Individual with State-Dependent Preferences

Notes: Baseline social preferences are the individual's nonmaterial motivations to contribute to the public good in the absence of an incentive. Incentives reduce the net cost of contributing to the public good; but unless $\lambda_c = 0 = \lambda_m$ (separability) or $\lambda_0 = 0$ (no social preferences to crowd out) they also affect the motivational salience of the individual's social preferences.

benefits of contributing, that is, the right hand side of (4) which we denote, θ. Considering the case in which there initially is no incentive, the effect 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

$$(5) \quad \frac{\Delta\theta}{\Delta s} = 1 + \lambda_0\left(\frac{\lambda_c}{\Delta s} + \lambda_m\right)$$

and is composed (as expected) of a direct effect (that is, 1, the top arrows in figure 1), and the indirect state-dependent effect that will be negative in the case of crowding out (λ_c or λ_m negative), and larger in absolute value the greater are the baseline

values of the individual (λ_0). We likewise see that

$$(6) \quad \frac{\Delta\theta}{\Delta\lambda_0} = 1 + \mathbf{1}\{s > 0\}\lambda_c + s\lambda_m,$$

which, in the case of crowding out, is declining in s.

Equations (5) and (6) make it clear that when λ_c and λ_m are negative, incentives and baseline values are substitutes: the effect of each on the marginal benefits of contributing varies inversely with the level of the other. The fact (from equation 5) that the crowding effect is larger for those with greater baseline social preferences makes sense and is consistent with experiments that have identified the strength of individuals' social preferences independently
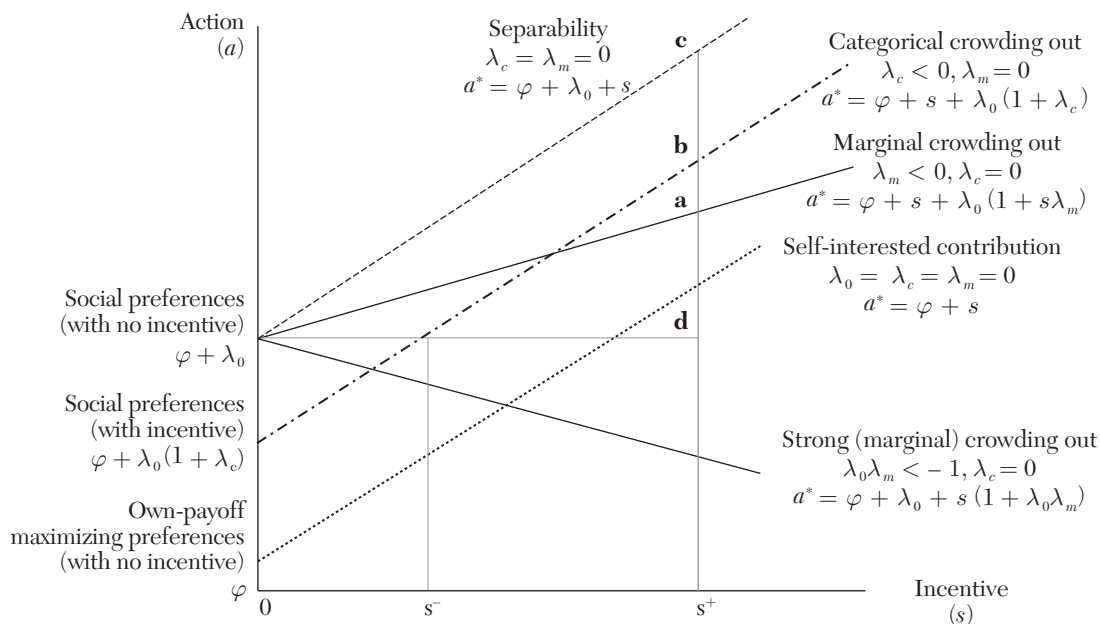
Figure 4. The Sophisticated Planner's Implementation Technology: Citizen's Contribution to the Public Good with State-Dependent Social Preferences

Notes: Under separability (top dashed line), incentives and social preferences are additive. Under strong marginal crowding out, the use of the incentive is counterproductive (i.e., reduces contributions). Under categorical crowding out (dot–dashed line), incentives are also counterproductive for a sufficiently small subsidy, namely, $s < s^-$.

of incentives and found that crowding out effects are larger for those with greater baseline values (Bohnet and Baytelman 2007; Kessler 2008; Carpenter and Myers 2010). This substitutability between incentives and baseline values will be important when we address questions of public policy in the penultimate section.

Using (5), we say that a particular change in incentives Δs has crowded out social preferences if $\Delta\theta/\Delta s < 1$, that is, if the total effect of the incentive is less than the direct effect, and conversely for the case of crowding in. Crowding will not occur if λ_c and λ_m or λ_0 are zero (that is, if social preferences are not state-dependent, or they are absent).

What we term strong crowding out holds if $\Delta\theta/\Delta s < 0$, which from the right-hand side of (5), we see may occur if categorical crowding out is large relative to the size of the subsidy, or its marginal effect ($\lambda_0\lambda_m$) is small, or if the marginal effect is negative.

The two forms of nonseparability are illustrated by crowding out in figure 4. Crowding in, which we do not show, would either shift the "separability" function upwards—categorical crowding in—or increase its slope—marginal crowding in. Because the functions in figure 4 represent the citizens' best responses to the planner's choice of an incentive and thus constitute one of the constraints making up the planner's optimizing

problem, we call these functions the planner's implementation technology.

An experiment allows an estimate of both categorical and marginal crowding out. Irlenbusch and Ruchala implemented a public goods experiment in which the 192 German student subjects faced three conditions: no incentives to contribute and a bonus, given to the highest contributing individual, that was either high or low (Irlenbusch and Ruchala 2008, details are in table 3 in section 5 and results are shown in figure 5). 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. Contributions in the low-bonus case were not significantly different from the no-bonus treatment. In the high-bonus case, significantly higher contributions occurred, but the amount contributed (53 units) barely (and insignificantly) exceeded that predicted for self-interested subjects (50 units).

In figure 5, we use 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 as in figure 4) to estimate the marginal effect of the bonus. We find 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 subjects without social preferences 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 (37.04) in the absence of any

incentive and the predicted contributions had an arbitrarily small incentive been in effect (the vertical intercept of the observed line in figure 5) 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 the Nash equilibrium for self-interested subjects, 12.04).

Categorical crowding out is also evident in other experiments. In one, 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 (Heyman and Ariely 2004). Using these data as we did in the Irlenbusch and Ruchala study, we estimate that the mere presence of the incentive reduced the willingness to help by 27 percent (compared to the no incentive condition).

Another experiment that allows us to distinguish categorical and marginal crowding was implemented by Cardenas (2004), but here (as in some other experiments) we observe categorical crowding in. Cardenas implemented an experimental Common Pool Resource Game very similar in structure to the kind of real world commons problem faced by his subjects—rural Colombian eco-system users. In the absence of any explicit incentives, the villagers on average extracted 44 percent less of the experimental "resource" than would have maximized their individual payoffs, providing evidence of a significant willingness to sacrifice individual gain so as to protect the resource and raise group-average payoffs. When they were liable to pay a small fine (imposed by the experimenter) if they overextracted the resource, as expected, they extracted even less than without the fine, showing that the fine had the intended effect.

The fact that the average extraction under the small fine treatment was 55 percent less than the Nash equilibrium for self-interested
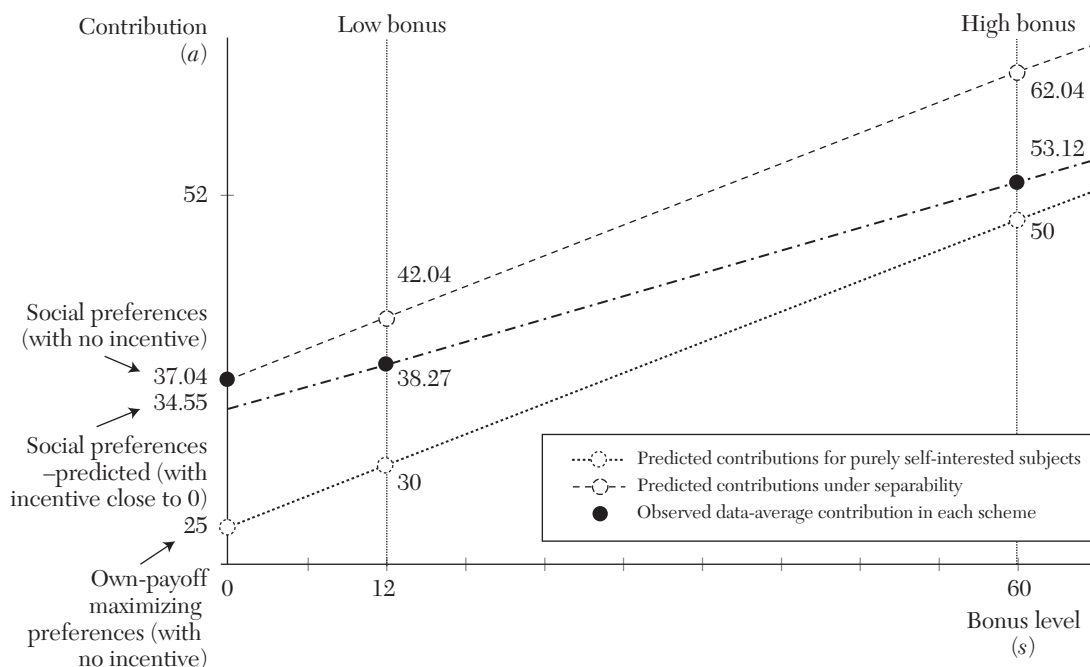
Figure 5. Categorical and Marginal Crowding Out

Source: Calculated from Irlenbusch and Ruchala (2008). See text. The experimental design is an adapted Public Goods Game comparing two team-based compensation schemes without and with a relative bonus.

subjects (when account is taken of the fine) suggests that the fine had increased the salience of the villagers' social preferences (by 25 percent, if the 44 percent deviation from the self-interested Nash behavior is taken as the measure of social preferences). Interestingly, raising the fine from a low to a high level had virtually no effect. Variations in the fine thus did not work as an incentive, but rather (in Cardenas's view) the very presence of the fine (high or low) was a signal, one that alerted subjects to the public good nature of the interaction. We will present other examples of fines as signals (section 3) and crowding in (section 9). These cases hold important lessons for why incentives sometimes are counterproductive and how

well-designed policies can make incentives and social preferences complements rather than substitutes.

Unfortunately, unlike the Irlenbusch and Ruchala and Cardenas studies, many experiments do not establish the response to incentives that would be observed under separability, so it is impossible to determine if incentives are "underperforming." A common misinterpretation of experimental results is to infer from the observation that an incentive has an effect in the intended direction that crowding out has not occurred (Rigdon 2009). But observing a positive incentive effect in an experiment does not preclude crowding out. It is clear from figure 4 (or equation 5 and the definition of

crowding out) that a positive incentive effect may occur in the presence of marginal crowding out (as long as it is not "strong") and in the presence of categorical crowding out (as long as the incentive is sufficiently large). For example, consider some substantial incentive indicated by s^+ in figure 4. Under both marginal and categorical crowding out, the action taken (points **a** and **b** respectively) is greater than in the absence of the incentive (**d**), so the incentive "worked": it affected the action in the intended direction. But the diagnostic for the presence of crowding is a comparison of these two action levels with the level that would have occurred under separability, namely point **c**, and this comparison makes it clear that crowding out occurred.

4. *Incentives Alter Cultural Learning: A Model of Endogenous Preferences*

A quite different mechanism by which crowding might occur has also been studied, one in which preferences are endogenous so that one or more of the parameters of the individual's value function—λ_0, λ_c and λ_m—are altered by incentives (Bar-Gill and Fershtman 2005; Hwang and Bowles 2011b). Hwang and Bowles present a model of cultural evolution in which the presence or level of incentives affects the process by which preferences are acquired or abandoned, so that a population's equilibrium distribution of preferences depends on incentives. By equilibrium preferences they mean a configuration of incentives and preferences such that the latter are stationary given the process of preference-updating.

In the Hwang and Bowles model, preferences are endogenous because (i) schools, families, religious organizations, and other societal institutions seek to promote civic minded values and (ii) individuals periodically alter their preferences in response to their own recent experiences. Their model of endogenous preferences is based on

two empirical regularities. The first is the powerful effect of mere exposure on preferences, documented by the social psychologist Zajonc (1968) and in subsequent works (Birch and Marlin 1982; Murphy and Zajonc 1993; Murphy, Monahan, and Zajonc 1995). The exposure effect is one of the reasons that cultural transmission may have what is termed a conformist component that favors the numerous over the rare, independently of their economic success. (See Boyd and Richerson 1985, 223ff; Ross and Nisbett 1991, 30ff; Bowles 1998 and the works cited there.) Following Boyd and Richerson, Hwang and Bowles assume a degree of conformist cultural transmission, so that the likelihood that an individual will adopt a particular preference varies not only with relative payoffs associated with the behaviors motivated by the preference but also with the prevalence of individuals with that preference in the population.

The second empirical regularity captured in their model of individual updating is that the presence and extent of incentives to contribute to a public project (or to engage in similar activities that benefit others) make the action (contribution) a less convincing signal of an individual's social preferences, resulting in observers interpreting some generous acts as merely self-interested. This is the key mechanism in the model of Benabou and Tirole (2006) showing how incentives may crowd out prosocial behavior. Similarly, in his "Generous actors, selfish actions" paper and in his subsequent work with Dufwenberg et al. (2011), Sobel (2009) and his coauthors provide not very restrictive conditions on individual utility functions such that "agents who care directly about the welfare and opportunities of others cannot be distinguished from selfish agents in market settings" (19). The reason is that for a class of utility functions admitting such other regarding preferences as inequality aversion, (paraphrasing the main theorem in their

2011 paper, 6) the "set of Walrasian equilibria of an economy [with other regarding preferences] coincides with the set of Walrasian equilibria of its corresponding . . . economy [in which] agents care only about their own direct consumption." Thus the use of market-like incentives may make it impossible to infer generous or fair-minded behaviors from the observed actions of one's fellow citizens.

There are two reasons why the presence of an incentive may lead people to mistake a generous act—helping another at a cost to oneself—for a self-interested one. The first is that the incentive provides a competing explanation of the generous act: "he did it for the money." The second is that incentives often induce individuals to shift from an ethical to a payoff maximizing frame (even relocating the neural activity to different regions of the brain); and knowing this, the presence of an incentive for an individual to help another may suggest to an observer that the action was self-interested (Gneezy and Rustichini 2000b; Heyman and Ariely 2004; Irlenbusch and Sliwka 2005; Li et al. 2009). The first "he did it for the money" reason depends on the magnitude of the incentive because in order to provide a convincing self-interested interpretation for the helping act the subsidy would have to exceed the cost of helping. The second reason— "when incentives are in force, everyone maximizes their payoffs"—is categorical; it is simply the presence of the incentive that matters. Of course, an observer could make the opposite mistake, inferring that the generous act that was motivated entirely by an incentive, was done for ethical rather than payoff maximizing reasons. In the model that follows the incentive is assumed on balance to degrade helping as the signal of a generous individual's type rather than motivating self-interested individuals to act in ways that are mistakenly taken as signals of a generous type.

Taken together, these two assumptions imply that the extensive use of incentives may reduce the perceived population frequency of individuals with social preferences, leading (via the conformist learning effect) to an evolutionary disadvantage of generosity over self-interest in the preference-updating process. To show this, Hwang and Bowles (2011b) adapt the model of endogenous preferences in Bowles (1998) and (2004) to study the effects of incentives on the preference-updating process. In terms of the state-dependent model of the previous section, they study the effect of incentives on the equilibrium fraction of the population for whom λ_0 is positive and sufficient to motivate contribution to a public good. The causal structure of their model is shown in figure 6.

Suppose there are two types: a Civic gives to the public good at a personal cost equal to $g > 0$ that may be partially offset by a subsidy s, while *Homo economicus* does not contribute and receives no subsidy. Both types update their traits by myopic best response, observing the material payoffs and public goods contribution of a sample of the population (they do not observe the utility of others) and a signal $\tilde{p}$ (possibly inaccurate when the planner implements a subsidy, $s > 0$) of p, the true frequency of the Civics in the population, so p, $\tilde{p} = \tilde{p}(s, p)$, which is decreasing in s. To capture the fact that the effect of the incentive on citizen's perception of the fraction of their fellow citizens who are Civics may depend on the mere presence of the incentive or on its extent (or both), let

$$(7) \quad \tilde{p} = p(1 + \mathbf{1}\{s > 0\}\Lambda_c + s\Lambda_m),$$

where as before the indicator $\mathbf{1}\{s > 0\} = 1$ if $s > 0$ and zero otherwise and $\Lambda_c \leq 0$ measures the categorical effect of the presence of an incentive on one's inference about another individual's type based on observing his or her contribution to the public good and $\Lambda_m \leq 0$ measures the marginal effect of
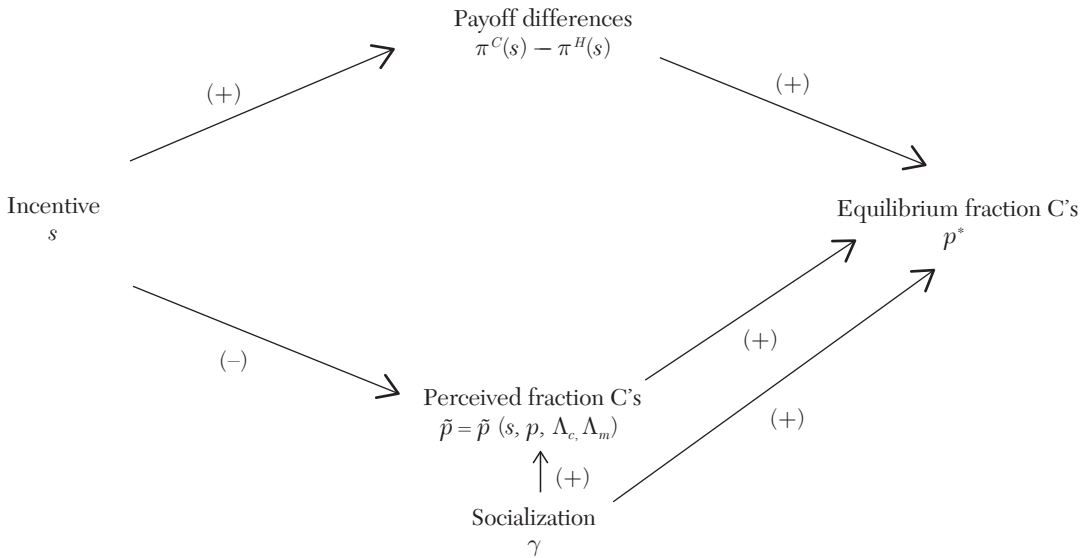
Payoff differences
$\pi^C(s) - \pi^H(s)$

$(+)$

$(+)$

Incentive
s

Equilibrium fraction C's
p^*

$(-)$

$(+)$

Perceived fraction C's
$\tilde{p} = \tilde{p}\,(s, p, \Lambda_{c,} \Lambda_m)$

$(+)$

$\uparrow (+)$

Socialization
γ

Figure 6. Crowding Effects of Incentives for a Population with Endogenous Preferences

Notes: Incentives raise the relative payoff of those contributing to the public good, supporting a larger fraction of civic-minded citizens; but unless $\Lambda_c = 0 = \Lambda_m$ (separability) or $p = 0$ in which case $p^* = 0$ (no civic minded citizens in the absence of a subsidy) a subsidy also alters the preference updating process by reducing the perceived fraction of civic-minded citizens $\tilde{p}$.

the level of an incentive on one's inference. Note that, when $p = 0$ or $s = 0$, $\tilde{p} = p$ so in the absence of the subsidy or when Civics are absent, the citizen's perception of the fraction of the population who are Civics is accurate.

The incentive has two offsetting effects on the distribution of types in the population, one intended and the other not: it raises the relative payoffs of the Civics, but it also reduces their apparent prevalence in the population. To see how this affects the equilibrium distribution of types in the population, suppose that individuals live forever but they periodically may switch their type. Denote the cultural fitness of trait i as $r^i(i = C, H$ for Civic and *Homo economicus*) defined as the expected number of replicas

that each individual bearing the trait will leave in the subsequent period. (If person k switches to j's type and j does not switch then k has left no replica and j has two replicas.) To capture the effect of socialization institutions on the evolution of preferences in this population, the authors suppose that in any period some fraction γ of the H-types will be converted to a C-type. (Because it plays little role in what follows, Hwang and Bowles do not model the manner in which socialization institutions accomplish this, other than to assume that the process is not affected by the level of incentives). They define $\alpha \in (0, 1]$ as the relative weight of conformism rather than payoffs in the updating process, β as the weight of payoff differences relative to the socialization effect and π_C, π_H as the

expected payoffs of the two types. Then, taking account of the common influences on the number of replicas, r_0, the cultural fitness of the two traits can be written:

$$(8) \quad r^C = r_0 + \alpha\left(\tilde{p} - \frac{1}{2}\right)$$

$$+ (1 - \alpha)\left[\beta(\pi_C - \pi_H) + \gamma\,\frac{1-p}{p}\right] \text{ and}$$

$$(9) \quad r^H = r_0 + \alpha\left(\frac{1}{2} - \tilde{p}\right)$$

$$+ (1 - \alpha)[\beta(\pi_H - \pi_C) - \gamma].$$

The second term in both equations is the conformism effect, and it favors the Civics if it is perceived that they constitute more than half of the population. The third term is the net effect of socialization and payoff based updating. The socialization effect in equation (8) (the second term in the square brackets) is derived as follows: noting that poulation size is normalized to unity, each of the $1 - p$ H types in the population has a γ probability of converting to C (shown in (9)) and thus appearing as $\gamma(1 - p)$ replicas assigned to the p Cs in the population. The final expression in (8) is thus the per C share of these socialized former Hs.

From these cultural fitness equations, one readily derives the familiar replicator equation for the movement of p over time:

$$(10) \quad dp/dt = p(1 - p)(r^C - r^H).$$

Taking account of the costs of contributing to the public good and the subsidy, and noting that the payoff difference between the types $\pi_H - \pi_C$ is just $g - s$, the resulting stationary condition for an interior value of p (namely $r^C - r^H = 0$, so that $dp/dt = 0$) is

$$(11) \quad \left(\tilde{p} - \frac{1}{2}\right)\frac{\alpha}{1 - \alpha} = \beta(g - s) - \frac{\gamma}{2p},$$

which requires that the conformist effect favoring the more common trait (the left-hand side) offset the net effect of that trait's payoff disadvantages and the societal level socialization effects (the right-hand side). Values of p satisfying (11) are termed the population's equilibrium preferences and denoted as $p^*(s)$.

Figure 7 illustrates the cultural equilibrium condition (11). The solid lines show the two sides of equation (11)—the conformist effect and the payoff plus socialization effects—and their intersection, satisfying equation (11) when $s = 0$ and giving $p^*(0)$ that is, the equilibrium distribution of preferences in the absence of incentives. The dotted lines show the effect of the implementation of a subsidy. The intended effect is to reduce the payoff advantage of the H types (they do not receive the subsidy) shifting downward the payoff cum socialization function. A naive social planner, unaware of the conformist effect would thus expect the introduction of the incentive to increase the fraction of Cs in the population to $p^N(s)$.

But the unintended effect of the subsidy is to reduce the perceived fraction of the population who are Cs and thereby to diminish the conformist advantage of the Cs. The downward shift in the conformist effect function thus partially offsets the payoff effect, with the resulting stationary distribution equal to $p^*(s)$. In the case of strong crowding out (not shown), the second effect would more than offset the first, resulting in a $p^*(s) < p^*(0)$.

The source of the nonseparability between socialization and incentives is clear if we return to equation (11) and consider the effect of an increase in s on the cultural fitness of the C types relative to the H types, evaluated at the status quo distribution of types in the population. In the initial absence of a subsidy, this is just the vertical distance at p^* (0) between the two functions that have been displaced by the introduction of the incentive (the dashed lines). Because this effect is
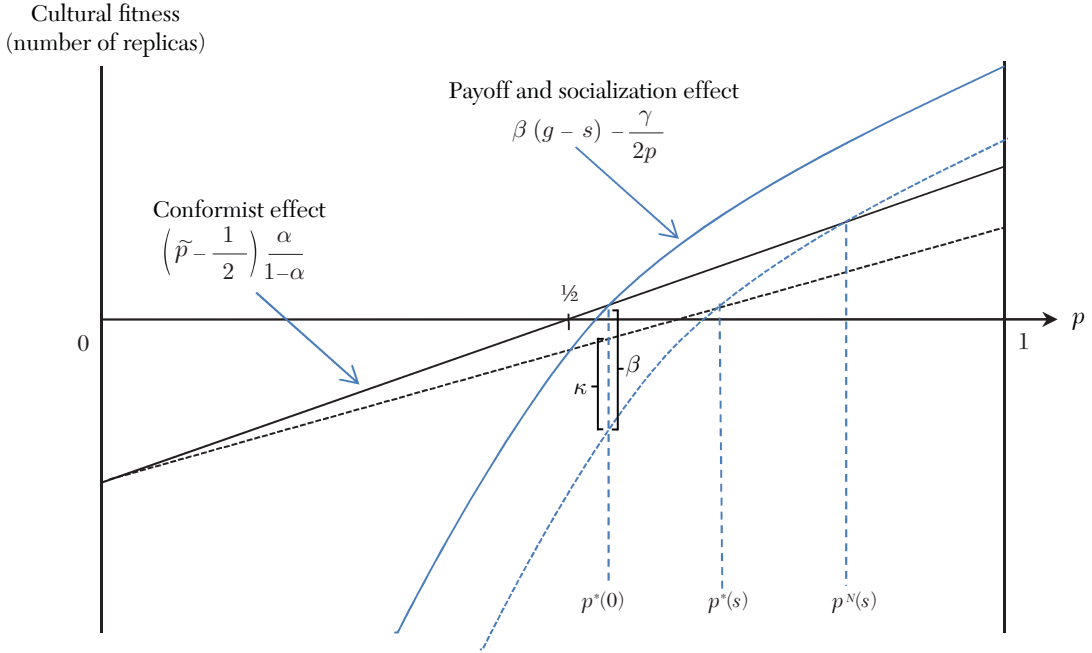
Cultural fitness
(number of replicas)

Figure 7. Incentives and Equilibrium Preferences

Notes: The figure (solid lines) shows the determination of fraction of citizens with social preferences in a cultural equilibrium under the influence of payoff-based and conformist updating, namely p^* when $s = 0$. The subsidy (dotted lines) reduces the payoff difference between *Homo economicus* and Civics, and in the absence of the effect on the perceived frequency of conformism in the population, the fraction of Cs in the population would increase from $p^*(0)$ to $p^N(s)$. However, the associated reduction in the conformism effect partially offsets this. The resulting equilibrium outcome is $p^*(s)$. The distance indicated by β is the effect anticipated by the naive planner who assumes separability, while κ is the true effect, taking nonseparability into account.

Source: Hwang and Bowles (2011b).

the cultural fitness advantage of the C-types following the introduction of the incentive when $p = p^*(0)$, Hwang and Bowles term it the evolutionary impact of the incentive, denoted by κ (analogous to θ in the state-dependent model of the previous section, κ is the difference between the two sides of (11)). The direct effect of incentives on κ is just β, but as is clear from the following expression, there also is an indirect effect:

$$(12) \quad \left.\frac{\Delta\kappa}{\Delta s}\right|_{p^*} = \beta$$

$$+ \frac{\alpha}{1-\alpha} p^*(\gamma, s)\left[\Lambda_m + \frac{\Lambda_c}{\Delta s}\right],$$

where the left-hand side means the change in the equilibrium condition associated with the change in s, for the given level of p, namely p^*. (As in the model of state-dependent preferences, we consider discrete changes

here rather than simply differentiating (11), in this case because of the discontinuity of $\tilde{p}$ at $s = 0$ in the presence of categorical crowding.)

The indirect effect will be negative in the case of crowding out, so the total effect of the incentive is less than the direct effect. The absolute size of the indirect effect (the second term on the right of (12)) is (as expected) increasing in the extent of conformism in updating and in the (absolute magnitude of the) crowding parameters. Importantly, the negative indirect crowding out effect will be larger in absolute value, the greater is p^*. Because p^* varies positively with the socialization effect (γ), the total effect of the incentive is less, the more effective is a society's socialization institutions. The crowding effect will absent (separability will hold) if $\Lambda_m = \Lambda_c = 0$ in which case $\tilde{p} = p$, so there are no misperceptions of the fraction of Cs in the population, or $\gamma = 0$ in which case there are no Cs to misperceive as self-interested, or $\alpha = 0$ in which case there is no conformism in updating so the misperceptions induced by the incentives have no effect.

We also have that the evolutionary impact of socialization institutions is

$$(13) \qquad \frac{\Delta\kappa}{\Delta\gamma}\bigg|_{p^*} = \frac{1}{2p^*(\gamma, s)},$$

which diminishes with greater use of incentives because (in the absence of strong crowding out) incentives raise p^*.

Thus where crowding out occurs incentives and socialization institutions are substitutes in the sense that the marginal effect of one on the evolutionary advantages of the civic minded types diminishes with the level of the other. We will return to the property of incentives and socialization as substitutes and the possibility of making them complements when we consider the policy implications of these models and the data to follow.

A summary of the two sources of nonseparability—state-dependence and endogeneity of preferences—and the mechanisms involved is provided in table 2.

The design of effective incentives in cases where separability may not hold requires a better understanding of the cognitive or affective effects of incentives that explain the categorical and marginal crowding out effects observed in experiments. We turn in the next three sections to the mechanisms that make preferences incentive-state-dependent, resulting in crowding out effects before considering (in section 8) the evidence for the adverse effects of incentives on preference-updating. (We consider crowding in—the case where incentives and social preferences are complements—in section 9.)

5. *Bad News: Incentives Provide Information about the Principal*

Incentives are implemented for a purpose, and because the purpose is often evident to the target of the incentives, the target may also infer information about the person who designed the incentive, about his or her beliefs concerning the target, and the nature of the task to be done (Benabou and Tirole 2003; Fehr and Rockenbach 2003).

We will illustrate this incentives-as-information-about-the-incentive designer effect by the negative response to fines imposed by experimental "investors" and "trustees" in the Trust Game, a principal–agent experiment implemented by Fehr and Rockenbach.

German students in the role of "investor" were given the opportunity to transfer some amount to the other player, called the "trustee." This amount was then tripled by the experimenter. The trustee, knowing the investor's choice, could in turn "back-transfer" some (or all, or none) of this tripled amount, returning a benefit to the investor (Fehr and Rockenbach 2003). When the investor

TABLE 2
SEPARABILITY AND CROWDING WHEN SOCIAL PREFERENCES ARE STATE-DEPENDENT OR ENDOGENOUS

Preferences	State dependent	Endogenous
Exogenous determinant of social preferences	Individual baseline values λ_0	Population level socialization effect γ
Crowding mechanism	Salience of values $v = \lambda_0(1 + \mathbf{1}\{s > 0\}\lambda_c + s\lambda_m)$	Perceived fraction C's $\tilde{p} = p(1 + \mathbf{1}\{s > 0\}\Lambda_c + s\Lambda_m)$
Intended target of the incentive	Individual best response $a^*(s, g(a), \lambda_0, \lambda_c, \lambda_m)$	Fraction of Cs in population $p^*(s, g, \gamma, \Lambda_c \Lambda_m)$
Separability	$\lambda_0\left(\dfrac{\lambda_c}{\Delta s} + s\lambda_m\right) = 0$	$\dfrac{\alpha}{1-\alpha}\, p^*(\gamma, s)\left(\Lambda_m + \dfrac{\Lambda_c}{\Delta s}\right) = 0$
Sufficient Conditions: separability	$\lambda_0 = 0$ or $\lambda_c = \lambda_m = 0$	$\gamma = 0, \alpha = 0$ or $\Lambda_c = \Lambda_m = 0$
Necessary Conditions: crowding out (in)	$\lambda_0 > 0$, λ_c or $\lambda_m < (>)0$	$\gamma > 0, \alpha > 0$ Λ_c or $\Lambda_m < (>)0$

Notes: In both models, the citizens may bear a cost ($g(a)$ or g) in order to contribute to a public good where a subsidy, s, may partially offset the cost. In the endogenous preference model, those who contribute are Cs. Additional notation: λ_m, λ_c and Λ_m, Λ_c are the marginal and categorical crowding parameters (in the state-dependent and endogenous cases, respectively) and α is the relative importance of conformism in the endogenous preferences model.

transferred money to the trustee, he or she also specified a desired level of the back-transfer. The experimenters implemented an incentive condition in which the investor had the option of declaring that he would impose a fine if the trustee's back-transfer were less than the desired amount. The investor could also decline the use of the fine, the choice of using or declining the fine option being known to the trustee and taken prior to the trustee's decision. There was also a "trust" condition in which no such incentives were available to the investor.

Trustees reciprocated generous initial transfers by investors with greater back-transfers. But the use of the fine reduced return transfers conditional on the investor's transfer, while renouncing the use of the fine when it was available to the investor increased back-transfers. Only one-third of the investors renounced the fine when it was available; their payoffs were 50 percent greater than the investors who threatened use of the fines.

The proximate causes of the negative impact of incentives in this case are suggested by evidence on the neural responses of the trustees in another Trust Game experiment (Li et al. 2009). As in the Fehr and Rockenbach experiment, the investor's threat of sanctions negatively affected back-transfers by trustees. To identify the proximate causes of this result, Li and his coauthors used functional magnetic resonance imaging (fMRI) to compare the

activation of distinct brain regions of trustees when faced with an investor who had threatened to sanction the trustee for insufficient back-transfers and an investor who had not threatened a sanction. Sanction threats deactivated the Ventromedial Prefrontal Cortex (VMPFC), a brain area whose activation was greater in trustees who made larger back-transfers, as well as other brain areas thought to be involved in the processing of social rewards. The threat activated the parietal cortex, an area thought to be associated with cost–benefit analysis and other self-interested optimizing processes. The interpretation by Li and his coauthors is that the sanctions induced a "perception shift" favoring a more self-interested response.

The signaling interpretation of counterproductive incentives in the Trust Game suggested by Fehr and Rockenbach is that in the trust condition, or when the fine was renounced by the investor, a large initial transfer signaled that the investor trusted the trustee. The positive response to the investor's renunciation of the fine option is a categorical effect, analogous to the negative categorical effect of the use of incentives in the Irlenbusch and Ruchala experiment described above. The threat of the fine, however, conveyed a different message and diminished the trustee's reciprocity.

Similar cases of crowding out due to the "bad news" conveyed by the incentive are at work in experiments among student subject pools in Switzerland, the United States, Italy, France and Costa Rica (as well as Germany) and in a diverse set of games including Gift Exchange, Public Goods, and a charity giving setting similar to a Dictator Game. Costa Rican businessmen also responded negatively to the bad news that incentives conveyed. Table 3 summarizes experiments in which this incentives-as-signals effect appears to have been at work (in some cases along with other mechanisms, to which

we now turn [16, 17, 19, 24, 28]. Please note that the numbers in square brackets throughout the paper refer to the experiments as numbered in the tables). Crowding out as the result of the "bad news" mechanism may be prevalent in principal–agent settings and can be averted where the principal has a means of signaling trust or fairness [1–3]. Not surprisingly crowding out affects individuals who are intrinsically motivated or fair-minded [5–6]; for own payoff maximizers, it appears there is nothing to crowd out.

6. *Moral Disengagement: Incentives May Suggest Permissible Behavior*

In most situations, people look for clues of appropriate behavior, and incentives often provide them. In table 4, we survey experiments in which this framing effect appears to have been at work. These experiments differ from those in table 3, in which incentives were deployed by experimental subjects in the role of a principal interacting with an agent. Here incentives are implemented exogenously, that is by the experimenter, so that they provide no information about the intentions or beliefs of other experimental subjects. As can be seen from the table, incentives appear to affect moral disengagement not only among students but also (as we have seen) among poor Colombian villagers [13, 14] and top U.S. CEOs [16]. Moral disengagement was evident in the Ultimatum Game and the Common Pool Resource Game [11–14; 20] as well as in the games for which the bad news mechanism was at work [1, 3, 6, 7, 9, 10]. In addition, this mechanism may be clearly recognized in settings of 1-player games (i.e., Dictator Game or a performance Task) [30, 32].

Hoffman and her coauthors (1994) illustrated the framing power of names: generosity and fair-minded behavior were diminished

TABLE 3
BAD NEWS: INCENTIVES PROVIDE INFORMATION ABOUT THE PERSON WHO IMPLEMENTS THE INCENTIVE (I)

Citation	Subjects (number)	Games or activities	Institutional environments (treatments)	Results relevant to separability	Comment (quotes are from the cited paper)
[01] Fehr and Rockenbach (2003)	German students (238)	Trust Game	• Optional punishment as an incentive contract (i.e. a fine if less than the desired back-transfer amount is returned). The level of the fine is fixed by the experimenter and the only choice of the investor is whether to impose the fine or not	Trustee's back-transfers are lower when investors impose fines. Not using the punishment option when it is available results in larger back-transfers and a larger joint surplus.	Explicit incentives undermine altruistic cooperation and reciprocity; forgoing the punishment option is a signal of good will and trust. See Fehr and List (2004). Negative effects of use of the punishment option are greater when the investor demands a larger share of the joint surplus. Categorical crowding out when the investor chooses the fine. **F**
[02] Fehr and List (2004)	Costa Rican CEOs (126) and students (76)	Trust Game	• Optional punishment as an incentive contract (i.e., a fine if less than the desired back-transfer amount is returned)	CEO principals trust more and are more trustworthy than students and as a result they achieve allocations closer to the maximum surplus that could be generated by the two parties. Joint surplus is highest when the punishment option is available and not used and lowest if the punishment option is used.	Key to performance: "the psychological message... conveyed by incentives – whether . . . kind or hostile . . ." (p. 745). See Fehr and Rockenbach (2003).
[03] Borges and Irlenbusch (2007)	German students (179)	Buyer–Seller Game	• Three rights of withdrawal: none, voluntary offer of a right of withdrawal (with a return cost for the seller) and imposed • The right of withdrawal when imposed has a return cost for the buyer or not	When sellers voluntarily offer a withdrawal right, buyers make order decisions that are less harmful for the seller than if the withdrawal right is imposed on sellers exogenously.	"Buyers are more inclined to behave fairly toward the sellers if they have . . . granted the withdrawal right voluntarily than if it is constituted by law". (p. 17) [because it is] "perceived . . . as a generous act and they might feel inclined to reciprocate by not exploiting the seller . . ." (p. 12). **F**
[04] Fehr and Schmidt (2007)	German students (70)	Gift-Exchange Game	• Two internal forms of enforcement: The principal (employer) can choose to rely on - an announced unenforceable bonus contract - a combination of the bonus contract with a fine	Most principals do not use the fine. The joint surplus under the pure bonus contract is 20 percent greater than under the combined contract. Wages are 54 percent higher in the pure bonus contract. Profits are not significantly different in the two contracts.	"Explicit and implicit incentives are substitutes rather than complements" (p. 3). Agents perceive that principals who are less fair are more likely to choose a combined contract and less likely to pay the announced bonus. The effect of effort on the bonus paid is twice as great in the pure bonus case.

(continued)

TABLE 3
BAD NEWS: INCENTIVES PROVIDE INFORMATION ABOUT THE PERSON WHO IMPLEMENTS THE INCENTIVE (I)
(continued)

Citation	Subjects (number)	Games or activities	Institutional environments (treatments)	Results relevant to separability	Comment (quotes are from the cited paper)
[05] Fehr et al. (2007)	German students (130)	Gift-Exchange Game	• Three internal forms of enforcement: The principal can choose to rely on - a trust (pure fixed payment) contract, or a price deduction (i.e., fine) contract - a trust, a fine, or an unenforceable bonus contract • Different frames: employer–employee or buyer–seller	Under the unenforceable bonus contract subjects contribute more than the payoff maximizing Nash equilibrium, outperforming the enforceable incentive contract (fine). The results are the same independently of the framing.	"Bonus contracts that offer a voluntary and unenforceable bonus for satisfactory performance provide powerful incentives and are superior to explicit incentive contracts when there are some fair-minded players" (121).
[06] Dickenson and Villeval (2008)	French students (182)	Gift-Exchange Game with a computer task	• Stranger or Partner with communication • Employer payoffs dependent on employee effort (variable) or not	In the partner treatment, when employer payoffs depend on employee effort less monitoring induces substantially higher performance. Consistent with Frey (1993).	While intrinsic motivation is evident in subject behaviors, in the Partner relationship the effect of more monitoring appears to be a reciprocity based negative response to the principal's lack of trust or intent to benefit at the agent's expense. **F**, **S**
[07] Irlenbusch and Ruchala (2008)	German students (192)	Public Goods Game	• An external form of enforcement (i.e., experimenter imposed): Team-based compensation with and without a reward for the highest contributor in the team • The reward is a low or a high bonus. • Pure individual bonus without team-based compensation	High (but not low) bonuses increase average effort, and joint surplus increases significantly only if the bonus is high, but decreases over time. Only with the purely team-based compensation (no individual incentives) do agents contribute more than self-interest would motivate. Pure tournament incentives induce effort levels below the self-interested Nash equilibrium prediction.	Both categorical and marginal crowding out occur. The tournament structure reduces voluntary cooperation. **F** (See text, pp. 381–82)

(continued)

TABLE 3
BAD NEWS: INCENTIVES PROVIDE INFORMATION ABOUT THE PERSON WHO IMPLEMENTS THE INCENTIVE (I)
(continued)

Citation	Subjects (number)	Games or activities	Institutional environments (treatments)	Results relevant to separability	Comment (quotes are from the cited paper)
[08] Ariely, Bracha, and Meier (2009)	U.S. students (161)	Charity giving based on task performance	• An external form of enforcement: With monetary compensation or without • Donation choices are public or private • Different frames: "good" and "bad" charitable causes	In the public treatment subjects exert more effort for a good cause and effort is substantially lower in the incentive treatment. Monetary incentives increase effort in the private treatment.	The signaling value of giving is compromised by incentives. "Image motivation is crowded out by monetary incentives [that are] more likely to be counterproductive for public pro-social activities than for private ones." (p. 1) Categorical crowding out. See Tenbrunsel and Messick (1999), Mulder, van Dijk, De Cremer, et al. (2006).
[09] Stanca, Bruni, and Corazzini (2009)	Italian students (96)	Gift-Exchange Game	• In the first move, Information (player 1 knows there is a second move) or No Information (player 1 does not know there is a second move and hence thinks the game is a Dictator Game)	Second movers' amounts returned are more correlated with the first mover's amounts sent in the No Information treatment.	Reciprocity is stronger in response to actions that are perceived as driven by intrinsic motivation, than to be in response to actions that are perceived as extrinsically motivated. **F**
[10] Fehr and Gächter (2002b)*	Swiss students (182)	Gift-Exchange Game	• Three external forms of enforcement: A Trust (pure fixed wage) contract, a deduction (i.e., fine) contract, and bonus incentive contract	Incentives reduce agent's effort. If the incentive is framed as a price deduction the effort reduction is greater than where the incentive is framed as a bonus. Incentives reduce total surplus, increase principal's profits.	Effects of incentives are due to the perceived fairness, kindness and hostility of the principal's action. **F, S**

Notes: The bold entries in the comments column—**I, F, S, E,** and **C**—indicate that the experiment in question could also have been included in tables 3 (**I**nformation about the principal), 4 (**F**raming), 5 (**S**elf-determination), 6 (**E**ndogenous preferences), or 7 (**C**omplementary relations between incentives and social preferences). All the papers but those marked with an * are published or forthcoming in a publication. The entries for each table are organized as follows: First, those studies that are published in a journal, ordered by year and first author. Second, working papers, ordered by year and first author. In a multiperiod experiment, the "stranger" treatment is one in which the pairing of subjects randomly changes from period to period. The "partner" treatment is one in which the pairing of subjects continues from period to period.

TABLE 4
MORAL DISENGAGEMENT: INCENTIVES MAY SUGGEST PERMISSIBLE BEHAVIOR (F)

Citation	Subjects (number)	Games or activities	Institutional environments (treatments)	Results relevant to separability	Comment (quotes are from the cited paper)
[11] Hoffman et al. (1994)	U.S. students (270)	Ultimatum Game; Dictator Game	• Roles are assigned by contest (the right to be the Proposer is 'earned' or randomly assigned). • Different frame: "Exchange" game (between a "seller" and a "buyer") or no frame • Anonymity: Double blind or not	Offers are lower and fewer low offers are rejected in the "Exchange" frame or when the proposer earns the right to his role. Proposers accurately gauge willingness of responders to accept lower offers. Dictators send lower amounts in double blind.	Institutional cues affect behavior: with property rights (i.e. legitimate 'earning' right to be proposer), a market framing or total anonymity proposers and responders are more self-interested. **S**
[12] Schotter et al. (1996)	U.S. students (247)	Ultimatum Game	• Survival treatment (two-stage): subjects with higher payoffs "survive" to proceed to stage 2. • Non survival treatment (one stage): the proposer is randomly assigned • Contextual framing: a simultaneous move normal or a sequential extensive form game	Competitive threats to survival induce lower offers, and fewer rejections of low offers.	The context affects behavior: 'earning' right to be the first mover or threat to survival induces proposers to behave in a more self-interested manner.
[13] Cardenas, Stranlund, and Willis (2000)	Colombian forest area dwellers (112)	Common Pool Resource Game	• External enforcement device with a low probability inspection and a fine • Communication	Fines induce more self-interested behavior and are ineffective at reducing common pool overexploitation in the longer run. Socially optimal individual deviations from the self-interested Nash equilibrium behavior (and the implied foregone payoffs by subjects) are least under the fines.	Weakly (exogenously) enforced fines diminish socially motivated behavior. Fine appears to have induced a shift from moral to self-interested frame. See Tenbrunsel and Messick (1999).
[14] Cardenas (2004)	Colombian users of rural ecosystems (265)	Common Pool Resource Game	• Different levels of external enforcement (low and high fines) with announcement of socially optimal extraction level and without communication • Communication without fines and announcement.	Deviation from self-interested behavior is much greater under communication (no fine) than under either high or low fines without communication. The behavioral effect of high (compared to low) fines is less than 6 percent of the predicted effect assuming self-interested preferences.	Marginal crowding out. (See text and also table 7; where Categorical crowding in also occurs).

TABLE 4
MORAL DISENGAGEMENT: INCENTIVES MAY SUGGEST PERMISSIBLE BEHAVIOR (F) *(continued)*

Citation	Subjects (number)	Games or activities	Institutional environments (treatments)	Results relevant to separability	Comment (quotes are from the cited paper)
[15] Heyman and Ariely (2004)	240 U.S. students (150+90)	A computer task and a puzzle task	• Different forms of compensation (cash, candy or a cash-denominated amount of candy) • Different levels of monetary compensation (none, low, medium)	In both the cash and the candy conditions, effort increases when the compensation level increases from low to medium. In the no-compensation treatment, effort is higher than the low-compensation condition for both the cash and the cash in terms of candy conditions and is not different from low-compensation in the candy condition.	The level and form of compensation affect performance. "Monetary compensation may act as a strong signal invoking norms of money markets instead of social-market relations" (p. 6) Monetary incentives influence the ways in which tasks are framed and the motivation to engage in them. The type of market in which the exchange takes place influences the relationship between reward and motivation.
[16] Bohnet and Baytelman (2007)	Senior executives in U.S. (353)	Trust Game and a Dictator Game	• No communication, face-to-face pre-play communication or post-play communication • An external form of enforcement (Post-play monetary punishment or not) • Stranger and Partner	Repetition and communication increase amount transferred and back-transferred; the option of punishment for low back transfers reduces back-transfers of other regarding trustees (those who send more in the Dictator Game).	"The availability of punishment destroys intrinsic trust and lowers people's willingness to reward trust" (p. 1) **I**
[17] Houser et al. (2008)	U.S. students (532)	Gift-Exchange Game	• A form of enforcement (Punishment as an incentive contract (i.e., a fine)) • Intention treatment: Punishment is assigned exogenously or imposed by investors	When back-transfer requests are high in relation to the sanction's size, regardless of whether the request is fair and regardless of whether punishment is intentional, punishment incentives have detrimental effects on the amount returned.	"Subjects interpret punishment as the price for self-interested behavior and the price, regardless of whether it was intentionally imposed, is an excuse for selfishness" (p. 15) Categorical crowding out when the investor chooses the fine. See Fehr and Rockenbach (2003) and Mulder et al. (2006) **I**
[18] Mellstrom and Johannesson (2008)	Swedish students (262)	Subjects are offered the opportunity to take a health exam to become blood donors	• With and without a monetary compensation for becoming blood donors • To choose between a monetary compensation and donating the same amount to charity	The incentive reduces the supply of prospective blood donors from 52% to 30% among women. No effect among men. Allowing individuals to donate the payment to charity eliminates the negative effect of the monetary compensation.	The monetary incentive may make it more difficult to signal social preferences, diminishing the signaling value of contributing. Charity option facilitates signaling. Overjustification appears also to be involved. **S**

(continued)

TABLE 4
MORAL DISENGAGEMENT: INCENTIVES MAY SUGGEST PERMISSIBLE BEHAVIOR (F) *(continued)*

Citation	Subjects (number)	Games or activities	Institutional environments (treatments)	Results relevant to separability	Comment (quotes are from the cited paper)
[19] Li et al. (2009)	U.S. citizens (104)	Trust Game	• Optional punishment as an incentive contract (i.e., a monetary sanction if less than the desired backtransfer amount is returned)	Trustees reciprocate relatively less when facing sanction threats, and the presence of sanctions significantly reduces trustee's brain activities involved in social reward valuation (VMPFC, LOFC, and amygdala), while significantly increasing activities in parietal cortex previously implicated in economic decision making.	Monetary sanctions "encourage activity within neural networks associated with self-interested economic decision making while simultaneously mitigating activity in networks implicated in social reward evaluation and processing" (p. 3) **I**
[20] Henrich et al. (2010) and Barr et al. (2009)* and personal communication from Barr and Henrich (March 2009)	15 societies Including U.S. students, African workers, Amazonian, Arctic, and African Hunter-gatherers. (428)	Dictator Game, Ultimatum Game, and Third-Party Punishment Game (TPG)	• Differences between societies • Subjects played in the following sequence keeping their role (active or passive): first DG, then the UG and finally the TPG (an explicit incentive, i.e., fine)	In many populations in the TPG the incentives provided by the fine do not induce higher offers, but rather have the opposite effect; factors that may influence self-interest calculations (i.e. wealth, income and household size) are significant predictors of allocations in the TPG (but not in the DG). Membership in a "world religion" positively associated with offers in the DG but not in the TPG.	The presence of the fine in the TPG appears to have reduced the salience of moral reasoning (derived from the teachings of the world religions) and enhanced subjects concerns with their own economic needs. (See text, pp. 397–98.)

Notes: The bold entries in the comments column—**I, F, S, E,** and **C**—indicate that the experiment in question could also have been included in tables 3 (**I**nformation about the principal), 4 (**F**raming), 5 (**S**elf-determination), 6 (**E**ndogenous preferences), or 7 (**C**omplementary relations between incentives and social preferences). All the papers but those marked with an * are published or forthcoming in a publication. The entries for each table are organized as follows: First, those studies that are published in a journal, ordered by year and first author. Second, working papers, ordered by year and first author. In a multiperiod experiment, the "stranger" treatment is one in which the pairing of subjects randomly changes from period to period. The "partner" treatment is one in which the pairing of subjects continues from period to period.

by simply relabeling an Ultimatum Game the "Exchange Game" and relabeling proposers and responders "sellers" and "buyers." The power of names has been confirmed in many (but not all) experiments since then (Zhong, Loewenstein, and Murnighan 2007) but in some cases (Ellingsen et al. 2011) the framing effect appears to have altered subjects beliefs about the actions of others rather than their preferences.

But literally naming the game is not necessary for framing effects to occur. Incentives alone may provide powerful frames for the decisionmaker. A year before the first reality TV Survivor show, Schotter and his coauthors found that market-like competition for "survival" among subjects reduced their concern for fairness in an Ultimatum Game experiment (Schotter, Weiss, and Zapater 1996). In this game, player 1 is given an endowment and asked to propose a division of it with player 2. Player 2, knowing the size of the endowment, decides whether to accept or reject the division. If player 2 accepts, then the proposed division is implemented. If player 2 rejects, both players receive zero. As is commonly observed in the Ultimatum Game, player 1 made quite generous offers and low offers were frequently rejected. But when the experimenters told the subjects that those with lower earnings would be excluded from a second round of the game, player 1 subjects offered less generous amounts to player 2, and player 2 accepted lower offers. The authors' interpretation was that: " . . . the competition inherent in markets . . . offers justifications for actions that, in isolation, would be unjustifiable" (38).

While plausible, direct evidence for this "moral disengagement" explanation is lacking because the social preferences that apparently accounted for fair behavior in the nonsurvival condition of the experiment were not measured. There are cases, however, in which the reduction in the salience of ethical reasoning induced by the presence of incentives can be detected.

A large team of anthropologists and economists implemented both Dictator and Third Party Punishment Games in fifteen societies ranging from Amazonian, Arctic and African hunter gatherers to manufacturing workers in Accra, Ghana, and U.S. undergraduates (Barr et al. 2009; Henrich et al. 2010). In the Dictator Game, an experimental subject is assigned a sum of money and asked to allocate some, all, or none of it to a passive recipient. The Third Party Punishment Game is a Dictator Game with an active onlooker (the third party) who observes the dictator's allocation. If the third party deems the dictator's allocation worthy of punishment he or she may then pay to impose a monetary fine on the dictator. One would expect that in the presence of a third party, the dictators would adjust their allocations upwards (compared to the two party standard Dictator Game) and thus avoid being fined. But this was not the case; fining was common; it occurred in 30 percent of the interactions across the study sites.

Surprisingly, in only two of the fifteen populations were the offers significantly higher in the Third Party Punishment Game than in the Dictator Game, and in four of the populations the allocations were significantly (and in some cases substantially) lower. In Accra, for example, where 41 percent of the dictator's allocations resulted in fines by the third party, the allocations were 30 percent lower in the Third Party Punishment Game than in the Dictator Game. The incentives provided by the fine did not induce higher allocations, but rather had the opposite effect. (For two groups, there was a significant positive effect of the fine option indicating that, in these two cases, the incentive had some effect; but as we have seen does not preclude crowding out.)

Crowding out of ethical motives is suggested by the fact that the dictator's adherence to one of the world's religions (Islam or Christianity, including Russian Orthodoxy) raised allocations in the Dictator Game by 23 percent (compared to those unaffiliated with a world religion). But in the Third Party Punishment Game, the estimated "religion effect" was reduced to just 7 percent of its value in the Dictator Game and it was not significantly different from zero. The presence of the incentive based on the fine appears to have defined the setting as one in

which the moral teachings of these religions were not relevant. Consistent with a crowding out interpretation of these results, the negative effect on the dictator's allocations of his or her economic need (number of children, conditional on a given level of income and wealth) was substantial (and statistically significant) in the Third Party Punishment Game, but in the Dictator Game this "economic need effect" was an order of magnitude smaller and not significantly different from zero.

In the Accra sample (Barr 2004), the dictator's allocation covaried significantly with the frequency of attendance at church or mosque in the standard two party Dictator Game; but this large "religion effect" vanished in the Third Party Punishment Game. The incentives implicit in the Third Party Punishment Game appear to have substituted economic motivations for moral concerns. These experiments are also consistent with our model of state-dependent preferences, in which crowding out operates via an effect of incentives on the behavior of those with preexisting social preferences.

7. Control Aversion: Incentives May Compromise Intrinsic Motives and Self-Determination

Recent experiments by economists surveyed in table 5 as well as nonexperimental studies in economics (surveyed in Frey and Jegen 2001) provide evidence for a third reason why social preferences may be state-dependent in ways leading to crowding out. Table 5 does not include the original "overjustification" experiments done by psychologists (referred to in the introduction). Unlike the experiments by psychologists where incentives are typically implemented by the experimenter, economists often model strategic interactions in which the same apparently control averse reaction occurs, so these experiments could also fall under the

"bad news" about the principal rubric presented in table 3 [6,10]. Moreover, framing effects may result in moral disengagement in some of these experiments [21, 26, 28, 31]. Crowding out effects of intrinsic motivation may be recognized in Ultimatum games [11, 12] and games where the experimenter is the principal [18, 28, 31]. We think it is likely that in these and other cases more than one mechanism is at work.

Falk and Kosfeld used a principal–agent game to explore the idea that "control aversion" based on the self-determination motive may be a reason why incentives sometimes degrade performance (Falk and Kosfeld 2006). Experimental agents in a role similar to an employee chose a level of "production" that was costly to them and beneficial to the principal (the employer). The agent's choice effectively determined the distribution of gains between the two, with the agent's maximum payoff occurring if he produced nothing. Before the agent's decision, the principal could elect to leave the choice of the level of production completely to the agent's discretion, or impose a lower bound on the agent's production (three bounds were varied by the experimenter across treatments, the principal's choice was simply whether or not to impose it). The principal could infer that a self-interested agent would perform at the lower bound or, in the absence of the bound, at zero, and thus imposition of the bound would maximize the principal's payoffs.

But in the experiment, agents provided a lower level of production when the principal imposed the bound. Apparently anticipating this response, fewer than a third of the principals opted for its imposition in the moderate or low-bound treatments. This minority of "untrusting" principals earned on average half of the profits of those who did not seek to control the agents' choice in the low-bound treatment, and a third less in the intermediate-bound condition.

TABLE 5
CONTROL AVERSION: INCENTIVES MAY COMPROMISE INTRINSIC MOTIVES AND SELF-DETERMINATION (S)

Citation	Subjects (number)	Games or activities	Institutional environments (treatments)	Results relevant to separability	Comment (quotes are from the cited paper)
[21] Gneezy and Rustichini (2000b)	Israeli students (160 for the main experiment)	50 IQ test questions (plus a Principal Agent Game)	• Different levels of monetary rewards for correct IQ test response (very low, low, high and none)	A discontinuity in the effect of incentives at zero. Small rewards degrade performance; large rewards enhance it.	The presence of the incentive substitutes extrinsic for intrinsic motivation. Categorical crowding out. See Gneezy (2003) **F**
[22] Gneezy and Rustichini (2000b)	Israeli students (180)	Collected donations from households	• Different levels of monetary rewards for the voluntary work (low, high and none)	Discontinuity at zero. Performance with small rewards is lower than performance with high rewards and both are lower than performance with no rewards.	The presence of the incentive substitutes extrinsic for intrinsic motivation. Categorical crowding out. See Gneezy (2003)
[23] Rustrom (2002)	U.S. students (110)	Creative task ('tower of Hanoi')	• Two forms of external enforcement (a penalty or a reward) • Different levels of the external enforcement (none, weak, strong)	Penalties degrade performance; large rewards induce better performance than small (but no better than the noincentive treatment)	Explicit incentives have a detrimental effect on performance, but only in the case of penalties, not in the case of rewards. Penalties 'distract' subjects. Categorical crowding out.
[24] Falk and Kosfeld (2006)	Swiss students (804)	Gift-Exchange Game	• Different levels of a lower bound of performance selected by the experimenter (low, medium, and high) • The principal could choose whether to impose the minimum level or not • The principal chooses the agent's wage and whether to impose the bound	Most agents perform minimally (namely at the lower bound) in response to the principals' controlling decision. Majority of the principals anticipate this and do impose the bound, earning higher profits as a result.	Imposing a lower bound compromises subject's sense of autonomy and signals distrust and low expectations that diminish agents' reciprocity and good will towards the principal. Categorical crowding out. (See text) **I**
[25] Xiao and Houser (2011)	U.S. students (72)	Public Goods Game	• Exogenous punishment: None, private (only the punished subject knows when a round is monitored and the amount of the resulting punishment), public (all members of a group are told that information)	Private punishment induces lower levels of contribution than public punishment.	Weak incentives crowd out cooperation when implemented privately, but the same incentives when implemented publicly (but anonymously) promote cooperation.

(continued)

TABLE 5.
Control Aversion: Incentives May Compromise Intrinsic Motives and Self-Determination (S)
(continued)

Citation	Subjects (number)	Games or activities	Institutional environments (treatments)	Results relevant to separability	Comment (quotes are from the cited paper)
[26] Gneezy (2003)*	U.S. students (400)	Proposer-Responder Game	• The responder has three forms of enforcement (a punishment at a given cost, a reward at a given cost and nothing) • Different levels of the responder's enforcement (weak, strong)	Non-monotonic effects of explicit incentives (fines and rewards) on performance (a W-shaped function). Offers are highest with large incentives (fine and reward), and lowest with small incentives. The no incentive case, when proposers simply dictate allocation, is intermediate.	Extrinsic incentives undermine intrinsic motivation: a small fine or reward changes the mode of behavior from "moral" to "strategic". See Gneezy and Rustichini (2000a, 2000b) and Mulder et al. (2006). Categorical crowding out. **F**

Notes: The bold entries in the comments column—**I, F, S, E,** and **C**—indicate that the experiment in question could also have been included in tables 3 (**I**nformation about the principal), 4 (**F**raming), 5 (**S**elf-determination), 6 (**E**ndogenous preferences), or 7 (**C**omplementary relations between incentives and social preferences). All the papers but those marked with an * are published or forthcoming in a publication. The entries for each table are organized as follows: First, those studies that are published in a journal, ordered by year and first author. Second, working papers, ordered by year and first author. In a multiperiod experiment, the "stranger" treatment is one in which the pairing of subjects randomly changes from period to period. The "partner" treatment is one in which the pairing of subjects continues from period to period.

Control aversion and the desire for self-determination are not the only effects of the principal's seeking to bind the agent. As anticipated by our discussion of the information content of incentives above, the imposition of the minimum in this experiment gave the agents remarkably accurate information about the principals' beliefs about them. In post-play interviews, most agents agreed with the statement that the imposition of the lower bound was a signal of distrust; and the principals who imposed the bound in fact had substantially lower expectations of the agents. The untrusting principals' attempts to control the agents' choices induced over half of the agents (in all three treatments) to contribute minimally, thereby affirming the principals' pessimism. Depending on the distribution of principal's priors about the agents, a population with preferences similar to these experimental subjects could support both a trusting and an untrusting (Pareto-inefficient) equilibria. Thus results in the Falk and Kosfeld experiment appear to be the result of both compromised self-determination and negative information about the incentive designer.

8. *The Economy Produces People: Incentives Alter How New Preferences Are Learned*

As in the Hwang and Bowles model introduced in section 4, incentives may also affect long-term change in motivations because they alter key aspects of how we acquire our motivations, influencing both the range of alternative preferences

to which one is exposed and the economic rewards and social status of those with preferences different from one's own (Bisin and Verdier 2001; Bowles 2004; Bar-Gill and Fershtman 2005).

Experiments of at most a few hours duration are unlikely to uncover the causal mechanisms involved in this process of durable preference change. This is because adopting new preferences is often a slow process more akin to acquiring an accent than to choosing an action in a game. As in the Hwang and Bowles model above, the developmental processes involved typically include population-level effects such as conformism, schooling, religious instruction, and other forms of socialization that are not readily captured in experiments. Acquiring new preferences (like a new accent) often takes place early in life and the learning process is strongly attenuated thereafter.

However, historical, anthropological, social psychological and other data (surveyed in Bowles 1998) provide evidence for endogenous preferences, showing that economic structures affect parental child rearing values, personality traits rewarded by higher grades in school, and other developmental influences. Additional evidence that preferences are endogenous comes from the experimental studies of fifteen small scale societies with extraordinarily varied economic structures, ranging from farming to hunting and gathering. In these studies, cross subject pool comparisons showed a strong association between the nature of the diverse economic tasks required to secure a livelihood—participating in large cooperative hunting teams in contrast to solitary work in forest slash and burn horticulture, for example—and its members' experimentally measured generosity and fair-mindedness in the Ultimatum Game (Henrich et al. 2005, 2010).

Despite the limitations of experiments for the investigation of preference change, we survey in table 6 a number of experiments that are consistent with durable learning effects of incentives. (We have placed all of the experiments consistent with preference endogeneity in this table; of course many of them also provide evidence of the mechanisms we have identified as affecting state-dependent preferences.) We take as evidence for this the fact that the apparent effect of incentives on preferences persists even when, in later stages of an experiment, incentives are withdrawn, suggesting that the prevalence of social preferences in a population may depend on exposure to incentives in the past, as in the Hwang and Bowles model.

An example follows. In the public goods experiment designed by Falkinger et al. (2000), an incentive mechanism induced subjects to contribute almost exactly the amount predicted for a self-interested individual, while in the absence of the incentive subjects contributed significantly more than would have been optimal for a self-interested individual. But, consistent with a change in preferences due to exposure to incentives, in the absence of incentives, subjects who had previously experienced the incentive system contributed 26 percent less than those who had never experienced it.

While the cultural diversity and variety of games appearing in table 6 are substantial, and we think the preference learning effects that we have detected in these experiments are indeed at work, we do not yet have experiments capable of testing the mechanism underlying the models of the influence of incentives on the evolution of preferences proposed by Hwang and Bowles, Bar-Gill and Fersthman, and others.

9. *Crowding In*

In section 2, we identified a number of cases in which crowding in may occur. For

TABLE 6
THE ECONOMY PRODUCES PEOPLE: INCENTIVES ALTER HOW NEW PREFERENCES ARE LEARNED (E)

Citation	Subjects (number)	Games or activities	Institutional environments (treatments)	Results relevant to separability	Comment (quotes are from the cited paper)
[27] Falkinger, et al. (2000) and personal communication from Gächter 18 February 2008.	Swiss students (196)	Public Goods Game	Incentive compatible (Falkinger (1996)) mechanism and no mechanism; large and small group size; Interior and corner Nash equilibria.	Subjects implement the self-interested level of contribution under the mechanism, but contribute substantially more than the self-interested level in its absence (until late in the 20 period experiments) (e.g., figure 5). After experiencing the mechanism, subjects contribute 26 percent less when it is withdrawn than those who have not experienced it.	By rewarding contributions and penalizing shirkers the mechanism may have relieved subjects' sense of moral responsibility and legitimated the pursuit of self-interest. The effects persisted after the withdrawal of the mechanism. **F**
[28] Gneezy and Rustichini (2000a)	Parents from ten day care centers in Haifa, Israel		• An explicit enforcement (i.e., fine) is imposed for lateness in six of these centers.	Tardiness doubles in the six treatment centers and persists even after the fine is removed. No change in the four control centers.	The modest fine may have signaled 'how bad' lateness is and/or is perceived as a price of a service and displaces an ethical frame by a strategic one: "A fine is a price." **I**, **F**, **S**
[29] Bohnet, Frey, and Huck (2001)	U.S. students (154)	Contract Enforcement Game (finitely repeated)	• Different legal institutions (low, medium or high contract enforcement probability) • Low contract enforcement in the last rounds for all sessions.	The probability of enforcement and/or the cost of breach in the early rounds have a non-monotonic effect on contract performance in the later rounds: intermediate levels of contract enforcement decrease trustworthiness, low levels and high levels of legal contract enforcement increase trustworthiness.	"If there is enough time for the crowding dynamics to unfold, environments with low contract enforcement can produce outcomes as efficient as high levels of enforcement" (p. 141) "by affecting behavior, institutions affect preferences." (p. 142) **F**
[30] Meier (2007)	Swiss students (11,379)	Contributions to two funds to support financially needy other students.	• Matching donations: For a single semester subjects' contributions are not matched or matched • Matching donations at high or low rates. No matching in subsequent periods	Matching increases contributions when they are in force. But those who experience matching are substantially less likely to make a contribution to either fund in subsequent periods; average contributions show a small, insignificant negative net effect of the incentive.	The negative matching effect is probably not due to the information it conveys on the neediness of the funds (larger effect for the smaller matching rate) or to the subjects' desire to compensate for higher matching induced contributions in the treatment period. **F**

TABLE 6
THE ECONOMY PRODUCES PEOPLE: INCENTIVES ALTER HOW NEW PREFERENCES ARE LEARNED (E) *(continued)*

Citation	Subjects (number)	Games or activities	Institutional environments (treatments)	Results relevant to separability	Comment (quotes are from the cited paper)
[31] Reeson and Tisdell (2008)	Australian Students (98)	Public Goods Game	• Three external forms of enforcement: - moral suasion in the form of a single sentence to the effect that the payoff to all would be higher if all contributed (all periods); - a binding minimum contribution unexpectedly introduced during 4 periods and then removed - none	While the regulation is in place (during the middle stage) contributions are significantly higher than in the initial stage in which only suasion occurs. After the regulation is removed, contributions are 20 percent lower than in the initial stage. The suasion treatment dramatically increases voluntary contributions compared to a no suasion control.	Suasion enhances and imposed minimum contribution reduces other regarding preferences. Categorical crowding out. **F**, **S**
[32] Burks, et al. (2009)	Swiss (139) and U.S. (113) bike messengers	Sequential Prisoners' Dilemma Game	Messenger exposure to performance based pay in their work place or not	In a restricted sample unlikely to be affected by selection bias, second movers' exposure to performance pay is associated with between 12 and 15 percent greater likelihood of defection on a cooperative first mover.	The fact that the effects are from a game having no obvious connection with the job suggests that preferences learned under the incentive conditions of the work place are adopted outside the workplace.
[33] Irlenbusch and Sliwka (2005)*	German students (84)	Gift-Exchange Game	• Two internal forms of enforcement: The principal can choose - a trust (pure fixed wage) contract - compensation contract (i.e., a variable piece rate) • Two different sequences for the contracts	Incentives reduce cooperation (i.e. effort level) and the effect persists after the incentive is removed. Where principals are constrained to offer fixed wages, the effort levels of agents are considerably higher than when employers can choose an incentive contract.	Incentives (price rate) alter principals' and agents' perception of the situation: "they lead agents to adopt an individual maximization frame . . . rather than a cooperative frame," "agents have a stronger concern for the principal's wellbeing in the pure fixed wage setting." (p. 23) **F**

Citation	Subjects (number)	Games or activities	Institutional environments (treatments)	Results relevant to separability	Comment (quotes are from the cited paper)
[34] Herrmann and Orzen (2008)*	British students (116)	Tullock Rent-Seeking Game or individual choice task and then a Prisoners' Dilemma Game	• Two different sequences (strategic vs. individual): • First stage: the two-player Tullock Rent-Seeking Game (with a different subject) or an individual choice task (with the same incentives). Second stage: a Prisoners' Dilemma	Players cooperate more when they previously played an individual choice task than when the previous game is competitive –strategic, one (i.e. the Rent-Seeking Game) Cooperation and reciprocity rates decrease after subjects are exposed to rent-seeking competition.	Subjects may perceive the interaction in the rent-seeking contest as a negative one. ". . . an individual's attitude towards others undergoes changes between different types of situations because they evoke different contextual cues". (p. 3) "the experience of over-competitiveness in the contest game creates a disposition of rivalry in subjects that some cannot immediately "turn off" when the experiment ends" (p. 26)
[35] Gächter, Kessler, and Konigstein (2010)*	Swiss students (500)	Gift-Exchange Game	• Three external forms of enforcement: a Trust (pure fixed wage contract), a deduction (i.e. fine) contract and a bonus incentive contract • Stranger and Partner • Different sequences	Under incentive contracts agents choose a self-interested best reply (effort) i.e., there is no voluntary cooperation. Experiencing well-designed contracts reduces voluntary cooperation even after incentives are withdrawn.	Incentives may have a lasting negative effect on voluntary cooperation. **F**

Notes: The bold entries in the comments column—**I, F, S, E,** and **C**—indicate that the experiment in question could also have been included in tables 3 (**I**nformation about the principal), 4 (**F**raming), 5 (**S**elf-determination), 6 (**E**ndogenous preferences), or 7 (**C**omplementary relations between incentives and social preferences). All the papers but those marked with an * are published or forthcoming in a publication. The entries for each table are organized as follows: First, those studies that are published in a journal, ordered by year and first author. Second, working papers, ordered by year and first author. In a multiperiod experiment, the "stranger" treatment is one in which the pairing of subjects randomly changes from period to period. The "partner" treatment is one in which the pairing of subjects continues from period to period.

example, the incentive may provide good news about the principal or it may lead to moral engagement rather than its opposite. In table 7, we survey a number of studies that show this result. These experiments are of special interest to the social planner not only because they would ideally point the way to the design of policies which would make incentives and social preferences synergistic (that is complements) rather than substitutes, but also because it appears that crowding in occurs more often in games with more than three players (Public Goods [38, 41, 42, 44, 47–49] and Common Pool Resource [14, 43] games) a common characteristic of public policy settings. In the

TABLE 7
INCENTIVES CROWD IN SOCIAL PREFERENCES (C)

Citation	Subjects (number)	Games or activities	Institutional environments (treatments)	Results relevant to separability	Comment (quotes are from the cited paper)
[36] Falk, Gächter, and Kovacs (1999)	Hungarian students (126, 38)	Gift-Exchange Game	• Stranger and Partner • Two social approval treatments (face to face, social pressure)	Partner treatment increased effort levels; social pressure has little effect. Wage effort relationship (based on reciprocity) is steeper under partner than under stranger.	Repeated interactions provide powerful incentives while enhancing both intrinsic reciprocity motives and concerns for equitable shares (social pressure adds little).
[37] Gächter and Falk (2002)	Austrian students (116)	Gift-Exchange Game	• Stranger and Partner	With repetition, effort levels are higher than one shot interaction and some self-interested subjects act strategically as reciprocators and then choose the minimal effort level in the last period	Repeated interaction strengthens reciprocity norms and induces 'imitated' reciprocity. "The social norm of reciprocity and the repeated game incentives are complementary" (p. 18).
[38] Masclet et al. (2003)	U.S. (96) and French (44) students (140)	Public Goods Game	• Two forms of Punishment with different levels of disapproval (from 0 to 10 points received by a subject from any other agent): Monetary punishment (subjects can reduce the monetary payoff of others after observing their decisions) and nonmonetary punishment (subjects express disapproval of others' decisions with no effect on others' earnings) • Stranger and Partner • Three stages: In the first and third stages without the punishment. In the second stage, with punishment	Both monetary and non-monetary sanctions induce higher and similar levels of contributions. Individuals tend to make higher contributions relative to the preceding period the higher punishment they have received and the lower their contribution was relative to the group average. When the Punishment device is removed, subjects having been the target of previous monetary sanctions show higher contributions than those targeted by nonmonetary sanctions.	Cooperation can be enhanced by nonmonetary sanctions for reasons that are not strategic and may require repeated interaction. It appears that nonmonetary punishment, while not affecting the best response of a payoff maximizing subject, nonetheless raised contributions by enhancing the salience of social motives such as shame or external peer pressure. See Lopez et al. (2011).
[-]* Cardenas (2004)	Colombian users of rural ecosystems (265)	Common Pool Resource Game	• Different levels of external enforcement (low and high fines) with announcement of socially optimal extraction level and without communication • Communication without fines and announcement.	Deviation from the self-regarding Nash extraction level was 29% greater under the small fine than with no fine.	Individuals consider the norm of cooperation that is proposed externally [the announced optimal level] when extracting (p. 238). Categorical crowding in. (See text 381–82 and also table 3; where marginal crowding out also occurs.)

TABLE 7
INCENTIVES CROWD IN SOCIAL PREFERENCES (C) *(continued)*

Citation	Subjects (number)	Games or activities	Institutional environments (treatments)	Results relevant to separability	Comment (quotes are from the cited paper)
[39] Henrich et al. (2005)	Foragers, herders, others in 15 small-scale societies (1,128)	Ultimatum Game	• Differences between societies in the level of market integration and the potential payoffs to cooperation	Substantial cross cultural covariation between the degree of market integration (engagement in market exchange) and both average UG offers and (unpublished) the propensity to reject low offers.	Mutually beneficial interactions in market interactions with strangers may support the evolution of cultures of fair-mindedness toward strangers; "*doux commerce*"? Hirschman (1977). This study also presents evidence of incentives alter how new preferences are learned **E**
[40] Falk, Fehr, and Zehnder (2006)	Swiss students (240)	Labor Market Game (one employer, three workers)	• With and without a minimum wage. • Two different sequences	The introduction of a legal minimum wage affects workers' fairness preferences leading to a rise in their reservation wages (which persists even after the minimum wage has been removed).	"Minimum wages [may] affect [subjects'] fairness perceptions" (p. 1376) creating moral "entitlements." Obligations activate and or enhance social preferences. See Galbiati and Vertova (2008), Vertova and Galbiati (2010)
[41] Tyran and Feld (2006)	Swiss students (102)	Public Goods Game	Levels of sanctions: none, mild and severe • Enforcement: external or self-imposed (by referendum)	Experimenter imposed mild sanctions do not significantly affect average contributions to the public good. Compliance is much improved if mild law is endogenously chosen.	Experimenter imposed sanctions raised the expected cost of freeriding without affecting behavior; a possible explanation is that only referendum imposed sanctions conveyed a signal of moral disapproval by peers.
[42] Herrmann et al. (2008a)	16 student pools around the world (1120)	Public Goods Game (Partner)	• Monetary Costly Punishment	Cooperation is higher in the punishment condition. However, the average payoff with the punishment condition is lower than the average without punishment in many countries. Weak norms of civic cooperation and the weakness of the rule of law in a country are significant predictors of antisocial punishment (targeting high contributors), which reduces the net benefits to the group.	Punishment is socially beneficial only if complemented by strong social norms of cooperation with strangers so that peer punishment induces shame rather than resentment. The quality of the formal law enforcement institutions and informal sanctions are complements.

TABLE 7
INCENTIVES CROWD IN SOCIAL PREFERENCES (C) *(continued)*

Citation	Subjects (number)	Games or activities	Institutional environments (treatments)	Results relevant to separability	Comment (quotes are from the cited paper)
[43] Rodriguez-Sickert, Guzmán, and Cárdenas (2008)	Rural Colombians from 5 communities (128)	Common Pool Resource Game	• Three different forms of external enforcement (A fine regime imposed, a fine proposed to the players and rejected or accepted by them, none) • Different levels of external enforcement (low, and high) for the imposed fine	Under all treatments other than no fine, groups start at high levels of cooperation. Cooperation remains high only when a fine, be it high or low, is in force. If the players reject the fine, cooperation slowly unravels. Presence of low fines prevented unraveling of cooperation.	When fines are rejected, the implied affirmation of social norms may have temporarily increased cooperation; reciprocal preferences (anger at low contributors) may account for the subsequent erosion of cooperation. Small fines enhance unconditional cooperation by relieving cooperators of the need to retaliate against defectors by withdrawing their own cooperation.
[44] Carpenter et al. (2009)	U.S. students (172)	Public Goods Game	• Costly punishment: subjects can punish noncooperators at a cost to themselves • Different team's residual claim (marginal per capita return on the public good) • Different group size	Shirkers are punished by peers and they respond by contributing more, even in the last round unless the frequency of reciprocators is too low or the group is too large. High contributors who are punished subsequently contribute less. (Unpublished results not reported in paper).	Altruistically motivated mutual monitoring, by enhancing shame-induced cooperation, supports high levels of team performance. Synergistic effects of social preferences and peer-imposed incentives. This study also presents evidence of incentives alter how new preferences are learned **E**
[45] Carpenter and Myers (2010)	U.S. volunteer firefighters (217) and nonvolunteer community members (189)	Dictator Game	• Exogenous variation in the presence and level of small stipends paid to volunteer firefighters	Small monetary incentives increase turnout to fighting fires for firemen unconcerned about image but have no effect on image-concerned firemen (the estimated negative effect is not significant).	The effect of image concerns increases with the visibility of the activity (training is a less visible activity than fighting fires). For firefighters with image concerns the positive direct effect of small extrinsic incentives is offset by the negative indirect effect of incentives on the image-value of fighting fires.

(continued)

TABLE 7
Incentives Crowd in Social Preferences (C) *(continued)*

Citation	Subjects (number)	Games or activities	Institutional environments (treatments)	Results relevant to separability	Comment (quotes are from the cited paper)
[46] Gächter, Nosenzo, and Sefton (2010)	British students (84)	Gift-Exchange Game with 3-member firms (one employer and two employees)	• Employees move sequentially (Employee 1 has pay comparison information (i.e., information about what coworker earns) and Employee 2 additionally has effort comparison information (information about how coworker performs) • Employers can offer high wages to both employees, a high wage to Employee 1 only, a high wage to Employee 2 only and low wages to both	A homogeneous wage does not affect effort when an employee is matched with a coworker that provides less effort. Reciprocity toward the employer is more pronounced when the coworker is hard-working, as effort is strongly and positively related to own wage and when the employer pays unequal wages to the employees. Exposure to pay comparison information in isolation from effort comparison information does not appear to affect reciprocity toward employers.	Unequal wages conditional on worker type may induce high levels of reciprocity based effort; unconditional employer generosity fails to recognize the 'deserving' worker, and is not reciprocated. Incentives and social preferences as complements. Workers respond to employers' recognition of their work effort and hence deservingness, not to employer generosity.
[47] Lopez et al. (2011)	Colombian fishermen (180)	Public Goods Game	• Public reminder about benefits of cooperation plus 1/5 chance of receiving private reminder of the social losses resulting from the individual's non-cooperative behavior (Guilt); receiving public reminder of the social losses resulting from the individual's non-cooperative behavior (Shame), facing an external low penalty for not contributing to the public good, facing an external high penalty for not contributing.	Priming subjects to feel guilty about low contributions did not affect average contributions, but the random public revelation of one's contributions (inducing shame) substantially increased contributions. Experimenter's imposition of the fine further increased contributions but the level of the fine has no effect.	Results suggest the importance of moral framing and that the fine did not work as an incentive but rather as a signal highlighting the salience of the ethical dimension of the problem. Categorical crowding in.
[48] Vertova and Galbiati (2010)*	Italian students (210)	Public Goods Game (and a Lottery Game)	• Different levels of a stated (non-binding) obligation to contribute (zero, low and high) • A symmetric incentive structure (a level of contribution less (more) than the minimum contribution could be subject to a probabilistic penalty (reward)) of low and medium size	There is a positive effect of the obligation, which is greater when it is combined with a weak monetary incentive than when no incentives are offered. A stronger monetary incentive does not result in an increase in contributions. The strong monetary incentive also has no effect on behavior in the absence of the stated obligation.	Incentives not only influence material payoffs but also frame recommended high contributions as obligations. Incentives and obligations affect people's behaviors by activating values and/or coordinating individuals' beliefs. See also Galbiati and Vertova (2008).

(continued)

TABLE 7
INCENTIVES CROWD IN SOCIAL PREFERENCES (C) *(continued)*

Citation	Subjects (number)	Games or activities	Institutional environments (treatments)	Results relevant to separability	Comment (quotes are from the cited paper)
[49] Barr (2001)*	Zimbabwean villagers (602)	Public Goods Game	• Two external forms of nonmonetary punishment i)Public announcement: each player announces her level of contribution to everyone present in the session, ii) Subjects could make public verbal statements about each other's decisions: lighthearted criticism or the withholding of praise during informal gatherings	After the introduction of the public announcement and public criticism subjects contribute more.	The fact that nonmaterial punishment raises contributions suggests that it induces shame or other social emotions (the best response for a self-interested individual was unaffected). See Gächter and Fehr (1999) and Mulder, et al. (2006).
[50] Serra (2008)*	British students (180)	Bribery Game (public official–citizen)	• Three different forms of external enforcement (no monitoring; top-down auditing, and an accountability system which gives citizens the opportunity to report corrupt officials)	Under the accountability system, fewer officials engage in corruption. The presence of only top-down auditing did not affect the number of officers who demanded a bribe but induced corrupt officials to demand a higher bribe than no monitoring.	"Nonmonetary costs activated by the bottom-up component of the combined system had a significant impact on the public official's decision to engage in bribery" (p. 17)

Notes: The bold entries in the comments column—**I, F, S, E,** and **C**—indicate that the experiment in question could also have been included in tables 3 (**I**nformation about the principal), 4 (**F**raming), 5 (**S**elf-determination), 6 (**E**ndogenous preferences), or 7 (**C**omplementary relations between incentives and social preferences). All the papers but those marked with an * are published or forthcoming in a publication. The entries for each table are organized as follows: First, those studies that are published in a journal, ordered by year and first author. Second, working papers, ordered by year and first author. In a multiperiod experiment, the "stranger" treatment is one in which the pairing of subjects randomly changes from period to period. The "partner" treatment is one in which the pairing of subjects continues from period to period.

penultimate and final sections, we will return to these questions when we consider the policy implication of nonseparability.

Synergy between incentives and social preferences may explain why fines imposed on free riders by altruistic peers in a Public Goods Game induce higher levels of contribution in subsequent rounds of play (Fehr and Gächter 2000). Of course, crowding in need not have been involved; individuals might have simply best-responded to the anticipated loss in payoffs associated with low contributions. But more than this is at work. Consistent with the interpretation that incentives imposed by peers activate shame or other social preferences, purely verbal messages of

disapproval have a substantial positive effect on free riders' subsequent contributions (Barr 2001; Masclet et al. 2003). When those who have contributed more than others are punished (as sometimes occurs, Herrmann, Thöni, and Gächter 2008), they subsequently contribute less, and costly retaliatory punishment sometimes results (Bowles and Gintis 2006; Carpenter et al. 2009; Hopfensitz and Reuben 2009). This appears to occur because the targets of the punishment feel hostility rather than shame.

There are also other mechanisms at work. The incentives and constraints typical of the rule of law and other institutional designs that limit the most extreme forms of antisocial behavior and facilitate mutually beneficial interactions on a large scale may enhance the salience of social preferences by assuring people that those who conform to moral norms will not be exploited by their self-interested fellow citizens (Bowles 2011). This may explain the Hokkaido University subjects who cooperated more in a public goods experiment when assured that others who did not cooperate would be punished (Shinada and Yamagishi 2007) despite the fact that this had no effect on their own material incentives (those told this were not subject to the punishment). They apparently wanted to be cooperative but wished even more to avoid being exploited by defectors. According to this interpretation, the fine imposed by the experimenter on any free riding liberated the individual to act prosocially without fearing being exploited by less cooperative players. The respondents may have exhibited what Bohnet and her coauthors call "betrayal aversion," which was attenuated by knowing that betrayal would be punished by a third party (Bohnet et al. 2008).

Market interactions may also favor the endogenous evolution of social preferences. In two sets of experiments in small-scale societies in Africa, Asia, and Latin America (Henrich et al. 2005, 2010), individuals from the more market-integrated societies gave more in the Ultimatum Game. The authors conjecture that this may be due to the fact that more market exposed subjects had the experience of mutually beneficial exchanges with strangers, much like in the anonymous experimental settings. A very different piece of evidence consistent with this interpretation is that subjects who were exposed to unobtrusive priming with words relating to markets and exchange prior to playing a Trust Game were more likely to trust their partner than were subjects exposed to primes unrelated to markets (Al-Ubaydli et al. 2011).

A distinct mechanism underlying crowding in was apparently at work in a public goods experiment by Vertova and Galbiati (2010). Consistent with the Cardenas experiment described in section 2, they found that the effect of a stated (nonbinding) obligation to contribute a certain amount was greater when it was combined with a weak monetary incentive than when no incentives were offered. A stronger monetary incentive did not result in an increase in contributions. The strong monetary incentive also had no effect on behavior in the absence of the stated obligation. The authors' interpretation (like that of Cardenas) is that the explicit incentives enhanced the salience of the stated obligation (see also Galbiati and Vertova 2008).

10. *The Lab and the Street: Can One Generalize from Experimental Evidence?*

The experimental evidence for nonseparability would not be very interesting if it did not reflect real-life behavior. Testing for separability in natural settings is difficult, but generalizing directly from experiments even for phenomena much simpler than separability is a concern in any empirical study (Falk and Heckman 2009) and is

often unwarranted (Levitt and List 2007). Consider, for example, the Dictator Game: typically more than 60 percent of the dictators allocate a positive sum to the recipient, and the average given is about a fifth of the endowment. We would be sadly mistaken if we inferred from this that 60 percent of individuals would spontaneously transfer funds to an anonymous passerby, or that the same subjects would offer a fifth of the bills in their wallet to a person who is homeless asking for help. Another example: while prosocial behavior in an experiment by Benz and Meier (2008) was correlated with nonexperimental behavior, subjects who reported that they had never given to a charity allocated 65 percent of their endowment to a named charity in a lab experiment.

A possible explanation of these discrepancies between experimental and real-world behavior is that most individuals are strongly influenced by the cues of appropriate behavior offered by the situation in which an action is taken (Ross and Nisbett 1991), and there is no reason to think that experiments are an exception to this context-dependent aspect of individual behavior. External validity concerns arise from four aspects of human behavioral experiments that do not arise in most well-designed natural science experiments. First, experimental subjects typically know they are under an unknown researcher's microscope, possibly inducing different behaviors than would occur under total anonymity or under the scrutiny of neighbors, family, or workmates. Second, experimental interactions with other subjects are typically anonymous and without opportunities for ongoing face to face communication, unlike many social interactions of interest to economists and policymakers. Third, subject pools may be quite different from the real-world populations of interest, in part due to the process of recruitment and self-selection. Finally, many of the experiments that provide evidence for the

salience of social preferences are deliberately structured as strategic interactions like the Ultimatum or the Public Goods Game that give scope for ethical or other-regarding behavior that may be absent in competitive markets and other important real-world settings (Sobel 2010).

It is impossible to know whether these four aspects of behavioral experiments bias experimental results in ways relevant to the question of separability. For example, the fact that in most cases subjects are paid a "show up fee" to participate in an experiment might attract the more materially oriented who may be less motivated by social preferences subject to crowding out; or knowing that the topic of the experiment was cooperation the subjects might be atypically civic minded.

We can do more than speculate about these problems. Baran and her coauthors asked if University of Chicago Graduate School of Business students who were more reciprocal in the Trust Game (those who as trustees most generously reciprocated large transfers by the investor) were also those most likely to contribute to the University upon graduation. They were (Baran, Sapienza, and Zingales 2010). Fehr and Lorenz Goette found that in a group of bicycle messenger workers in Zurich, those who exhibited loss aversion in a laboratory experiment exploring the subjects' preferences over lotteries also exhibited loss aversion when faced with real-life wage rate changes (Fehr and Goette 2007). Karlan (2005) implemented a Trust Game among Peruvians participating in a micro-credit program; those who were least trustworthy (transferred less back to the "investor") in the experiment were less likely to repay their real world loans. Cohn and his coauthors (Cohn, Fehr, and Goette 2011) found that reciprocators in the lab (measured by play in a sequential PD Game) responded positively to a randomly awarded fixed wage increase in their work, while those who played the sequential PD

in a payoff maximizing way did not respond to the wage increase.

Among the Japanese shrimp fishermen that Carpenter and Seki (2010) studied, those who contributed more in a public goods experiment were more likely to be members of cooperatives that shared costs and catch among many boats than to fish under the usual private boat arrangements. A similar pattern was found among fishermen in the Brazilian north east, where some fish offshore in large crews whose success depends on cooperation and coordination, while those exploiting inland waters fish singly. The ocean fishers were significantly more generous (in Public Goods, Ultimatum, and Dictator Games) than the inland fishers (Leibbrandt, Gneezy, and List 2010).

A better test of the external validity of experiments would include a behavior-based measure of how cooperative the individuals were, not simply whether they took part in a cooperation-sensitive production process. The Brazilian fishers provide just such a test. Shrimp are caught in large plastic bucket-like contraptions; holes are cut in the bottom of the traps to allow the immature shrimp to escape, thereby preserving the stock for future catches. The fishermen thus face a real-world social dilemma: the present value of expected income of each would be greatest if they cut only small holes in their own traps while others cut large holes in theirs. Small trap holes are a form of defection, and just as in the Public Goods Game it is the dominant strategy for a self-interested individual. But a shrimper might resist the temptation to defect if he were both public spirited toward the other fishers and sufficiently patient to value the future opportunities that they all would lose were he to use traps with smaller holes. Fehr and Leibbrandt implemented both a Public Goods Game and an experimental measure of impatience with the shrimpers. They found that both patience and cooperativeness in the game predicted larger trap holes (Fehr and Leibbrandt 2011). The effects, controlling for a large number of other possible influences on hole size, were substantial. A shrimper whose experimentally measured patience and cooperativeness is a standard deviation greater than the mean is predicted to cut holes in his traps that are half a standard deviation larger than the mean.

Additional evidence of external validity comes from a set of experiments and field studies with 49 groups of herders of the Bale Oromo people in Ethiopia who were engaged in forest commons management. Rustagi and his coauthors implemented public goods experiments with a total of 679 herders. They also studied the success of the herders' cooperative forest projects. The most common behavioral type in the experiments, constituting a bit more than a third of the subjects, were "conditional cooperators" who responded positively to higher contributions by others. Controlling for a large number of other influences on the success of the forest projects, the authors found that groups with more conditional cooperators were more successful, in terms of number of new trees planted, than groups with fewer conditional cooperators. This was in part because members of groups with more conditional cooperators spent significantly more time monitoring the use of the forest by others. As in the case of the Brazilian shrimpers, the effects of group composition were large. A 10 percent increase in the fraction of experimentally identified conditional cooperators in a group was associated with an increase in trees planted or time spent monitoring by members of the group of about 3 percent (Rustagi, Engel, and Kosfeld 2010).

The available evidence suggests that students volunteering for experiments are not more pro-social in their orientations than

other students (Falk, Meier, and Zehnder 2011); nor are student subjects more prosocial than nonstudents, indeed the reverse seems to be the case. (Fehr and List 2004; List 2004; Cardenas 2005; Carpenter, Burks, and Verhoogen 2005; Bellemare, Kröger, and van Soest 2008; Carpenter, Connolly, and Myers 2008; Burks, Carpenter, and Goette 2009; Baran, Sapienza, and Zingales 2010; Cleave, Nikiforakis, and Slonim 2010; Cardenas 2011; Falk, Meier, and Zehnder 2011. For a description of these studies, see appendix A9 of Bowles and Gintis 2011.)

Levitt and List (2007) are right that care should be taken in generalizing experimental behavior to the real world. But none of the external validity concerns we have considered is sufficient to dismiss the experimental evidence that social preferences are important behavioral motivations and that these preferences may be affected by explicit incentives. This is especially the case when experimental subjects exhibit motives such as reciprocity, generosity and trust that allow a consistent explanation of otherwise anomalous real world examples of crowding in or out, such as those mentioned at the outset.

11. *Optimal Incentives for the Sophisticated Social Planner*

There are multiple plausible interpretations of the mechanisms underlying nonseparability in the experiments we have presented, as is clear from the substantial size of the intersections among the hypothesized crowding out mechanisms that is evident in figure 2. It would nonetheless be difficult, in light of these data, to sustain the implicit separability assumption adopted in many economic models.

A sophisticated social planner (or mechanism designer)—one who knows that the separability assumption is likely to be violated—faces a challenge that has yet to be addressed in the public economics literature: how to design optimal taxes, fines, or subsidies when the preferences that will determine citizens' responses depend on the incentives deployed. Thus, the designer must consider the effects—whether state-dependent or endogenous—of the instruments under consideration on individuals' social preferences and evaluate alternative policies on the basis of the resulting joint equilibrium of these preferences and economic allocations.

The problem facing the planner is quite a bit more difficult than the one we faced writing this paper. We studied the effects of incentives in experiments and natural settings and then sought ex post to determine the kinds of nonseparability—categorical or marginal crowding out or in—that might explain the results. The planner, however, must determine ex ante whether the separability assumption is likely to be violated, and if so, how. The challenge is even greater because the nature and extent of nonseparability itself is not given but (as we will see) may be influenced by the overall policy package of which the incentives are a part.

We begin with the more modest way of addressing the planner's problem and consider as exogenously given the nature and degree of the indirect effects of incentives on social preferences (that is, the signs and the size of the crowding parameters λ_c, λ_m, Λ_c, and Λ_m). Given the nature and extent of nonseparability, we then seek to determine the optimal level or mix of incentives taking account of their effects on preferences (Fershtman and Heifetz 2006; Heifetz, Segev, and Talley 2007; Bowles and Hwang 2008; Hwang and Bowles 2011a).

Here, two results may guide the social planner. The first is that, in the presence of crowding out, incentives and social preferences are substitutes, so the deleterious indirect effect of incentives will be least where

individual social preferences are modest or nonexistent (as will be the case in the endogenous preference model if there are few or no public spirited citizens or in the state dependent model where the citizen's baseline social preferences are modest or zero). Societies in which social preferences are more prevalent not only may be able to afford less use of incentives but will find them less effective (when both direct and indirect effects are accounted for) than would be the case in a less civic minded culture. By the symmetry of the definition of substitutes (see equations 6 and 13) in the presence of crowding out, policies to enhance social preferences (that is raising γ or λ_0) will be more effective in promoting contributions to the public good where incentives are little used.

In a cultural–institutional dynamic setting where economic incentives and socialization practices to promote civic mindedness are adopted as alternative measures to enhance public goods provision, this substitutability property of incentives and social preferences may support at least two evolutionarily stable equilibria. In one, extensive use of incentives is coupled with relatively low levels of civic mindedness in the population. In this state, there is little reason to inculcate social preferences, the effect of which would be modest given the crowding phenomenon resulting from the extensive use of incentives. In the other cultural–institutional equilibrium, a social planner serving a civic minded population makes more modest use of incentives due to their limited effectiveness, once their crowding out effects are accounted for.

The second result for the social planner takes us back to Titmuss and others who concluded that if incentives crowd out social preferences then incentives will be overused by a naïve planner who is unaware of the effects of incentives on preferences. As a result, in these cases the sophisticated planner would either not use incentives, or

would use them less than would the naïve planner. But the prescription that incentives are overused does not follow from the (correct) observation that crowding out occurs: it is readily shown that when crowding out occurs the sophisticated planner may make either greater or lesser use of explicit incentives than would her naïve counterpart (Bowles and Hwang 2008; Hwang and Bowles 2011a).

The reason why the planner may make greater use of incentives when incentives crowd out social preferences is that if incentives work less well than would be the case under separability, then there are two offsetting influences on their optimal use. The one that forms the basis of the Titmuss critique is that crowding out reduces the marginal effect of the subsidy on the target's behavior; and if this were the only effect Titmuss would be right. But there is a second often overlooked effect. Because the incentive is less effective (either categorically or marginally), the under provision of the public good will be exacerbated (compared to what would occur were crowding out absent) and if the benefits of the public good are concave in the amount provided, the marginal benefit of altering the target's behavior is therefore correspondingly greater.

The intuition is transparent: the doctor who discovers that a treatment he has been prescribing is less effective than he thought may opt for stronger doses rather than weaker or for abandoning the treatment. As long as there are diminishing marginal returns to the public good and crowding out is categorical (and not too large) the naïve social planner will make too little use of the incentive. The reason is that in this case crowding does not change the marginal effect of the incentive on the citizens' contribution level; but the reduction in the public good resulting from crowding means that the marginal benefits to increasing its supply rise. (If categorical crowding is sufficiently large, the naïve

planner will overuse the incentive because the sophisticated planner will choose no incentive at all in this case.) But the sophisticated planner may make greater use of incentives even when only marginal crowding out occurs, if the benefit function is sufficiently concave.

A less modest approach to the design of appropriate incentives where separability may not hold is to recognize that the extent of the nonseparability problem (that is, the magnitudes of the crowding parameters in the models of section 3 and 4, namely λ_c, λ_m, Λ_c, and Λ_m) is not exogenous, but can be affected by the nature of the incentives and the manner in which they are deployed. Designing policies that can convert incentives from being substitutes for social preferences to being their complements, however, requires an understanding of why crowding out occurs.

The most plausible explanation for the failure of the separability assumption is that when people engage in trade, produce goods and services, save, and invest, they are not only attempting to *get* things, they are also trying to *be* someone, both in their own eyes and in the eyes of others (Cooley 1902; Yeung and Martin 2003; Akerlof and Kranton 2010; Bloom 2010). We refer to the second—the being or becoming motives—as constitutive. Incentives addressed to our acquisitive desires sometimes appear to dampen or impede the pursuit of our constitutive aspirations. Among the reasons, we have seen, are that in addition to affecting the costs and benefits of an action, incentives also provide information about the person imposing the incentive, suggest appropriate behavior by framing decision situations, may compromise the target's sense of autonomy, and alter the environments in which we learn new preferences.

This may explain why incentives for settlement of conflicts may fail. Representative samples of Jewish West Bank settlers in 2005, Palestinian refugees in 2005, and Palestinian students in 2006 were asked how angry and disgusted they would feel or how supportive of violence they might be if their political leaders were to compromise on contested issues between the groups. Those who regarded their group's claims (on Jerusalem, for example) as reflecting "sacred values" (about half in each of the three groups) expressed far greater anger, disgust and support for violence if the compromise were accompanied by a monetary compensation for their own group than if no compensation were offered (Ginges et al. 2007). Similar results were fund in a survey of the willingness of Swiss citizens to accept environmental hazards (Frey and Oberholzer-Gee 1997). (For a discussion on environmental motivation and crowding effects, see Frey and Stutzer 2008.)

The importance of constitutive rather than acquisitive motives may be at work in the negative response to incentives that convey adverse information about the individual imposing the incentives. Recall that in the Trust Game implemented by Fehr and Rockenbach (2003) the investor's threat to fine the trustee if the back transfer was not sufficient resulted in a lesser level of reciprocity of the trustee: conditional on the investor's transfer to the trustee, back-transfers were less under the fine condition. This was especially the case when it appeared that the intent of the fine was to induce the trustee to grant most of the joint surplus to the investor. Where the announced desired back-transfer would have allowed the investor to capture most of the surplus had the trustee complied, use of the threat of the fine was very counterproductive.

The back-transfers from those facing an investor who declined use the fine averaged a generous 60 percent of what the trustee could have kept; while those threatened with a fine transferred just 22 percent (almost half gave nothing and paid the fine!). But where the investor announced a desired back

transfer such that the trustee who complied would retain half or more of the joint surplus, back transfers from those threatened with fine for noncompliance were slightly lower but not significantly so than among those facing an investor who chose not to use the fine. It appears that the use of the fine in these conditions signaled the unfair intent of the investor, rather than simply his distrust of the trustee.

Interpretations other than this "unfair intent" result are possible, however, for the larger was the desired back transfer, the more costly was compliance. Thus for larger demands, simply returning nothing and paying the fine (as many of the subjects did) might have been attractive to self-interested subjects, who had they been faced with a lower demand would have maximized payoffs by complying. It seems from this and similar experiments that fines may have negative effects even if imposed to implement a fair outcome and even when the decision to use the fine was not made by the investor, but rather by chance (Fehr and List 2004; Houser et al. 2008). But the experiments also are consistent with the idea that threats deployed in self-interested ways can backfire.

This helps explain the very different effect of incentives imposed by peers who do not stand to benefit personally. An example is the Public Goods experiment in which fellow group members have the opportunity to reduce their own payoffs in order to punish (reduce the payoffs of) others in their group once each member's contributions are revealed (Fehr and Gächter 2000, 2002a; Masclet et al. 2003). One treatment in these public goods experiments is particularly revealing: group membership is shuffled after each period so that in subsequent periods a punisher will not be in the same group with the target of his or her punishment, and thus cannot benefit from the target's response. Punishment in this case is an altruistic

act as it benefits others at the expense of the punisher and hence it cannot be interpreted as a signal of unfair intent. In this setting there is a strong positive response by low contributors who are punished.

Although there is no direct evidence, a plausible explanation of the effectiveness of incentives in this case is that, when punished by a peer who had nothing to gain by doing so, those who have contributed less than others interpret the punishment as a signal of public-spirited social disapproval by fellow group members seeking to uphold a social norm and willing to sacrifice payoffs to do so. As a result, targeted free riders and even free riders who escaped punishment feel shame, which they redress by subsequently contributing more. In this case the incentive (prospect of peer imposed fines) has crowded in social preferences. These experiments illustrate the opposite of the "bad news about the principal" results in section 5. The principals here are the peers who punish free riding fellow group members; and the positive response to the fines in this case may reflect the fact that the willingness to pay to punish defectors with no expectation of personal gain is good news about the person implementing the incentive.

Consistent with the interpretation that crowding out does not follow from the use of incentives per se but rather from the meaning that the incentives convey to the participants, is an extension of the "control aversion" experiment of Falk and Kosfeld (2006) described in section 5. Schnedler and Vadovic (2011) found that when agents themselves implemented controls (rather than the principal) the negative response did not occur. A large number of experiments have found positive effects of incentives imposed by the decision of the targets of the incentives rather than by the experimenter or by a principal (Cardenas 2005; Tyran and Feld 2006; Kroll, Cherry, and Shogren 2007; Ertan, Page, and Putterman 2009; Kosfeld,

Okada, and Riedl 2009; Mellizo, Carpenter, and Matthews 2011; Sutter, Haigner, and Kocher 2011).

John Stuart Mill (whose definition of the restrictive boundaries of our discipline we mentioned at the outset) and economists since have recognized that the purposes of individual economic action are constitutive as well as acquisitive (Akerlof and Kranton 2010). But what some have missed is that our acquisitive and constitutive motivations may not be separable.

Some of the founders of economics knew this. Jeremy Bentham's *Introduction to the Principles of Morals and Legislation* (1789), is arguably the first text in what we now call public economics. In it he explained how proper incentives should harness self-interested objectives for public ends by making "it each man's *interest* to observe . . . that conduct which it is his *duty* to observe." In other words, make sure that doing his duty is incentive compatible. But he also understood the constitutive side of action and the need to design incentives that are complements of the moral sentiments rather than substitutes: "A punishment may be said to be . . . a moral lesson, when by reason of the ignominy it stamps upon the offence, it is calculated to inspire the public with sentiments of aversion towards those pernicious habits and dispositions with which the offence appears to be connected; and thereby to inculcate the opposite beneficial habits and dispositions" (Bentham 1970 [1789], 26).

Few economists followed Bentham in this. An exception is Albert Hirschman, who noted that economists seek "to deal with unethical or antisocial behavior by raising the cost of that behavior rather than proclaiming standards and imposing prohibitions and sanctions. The reason is probably that they think of citizens as consumers with unchanging or arbitrarily changing tastes in matters civic as well as commodity-related behavior. . . A principal purpose of publicly proclaimed laws and regulations is to stigmatize antisocial behavior and thereby to influence citizens' values and behavioral codes (Hirschman 1985, 10).

The fact that punishments are not only incentives but also "moral lessons" that "stigmatize antisocial behavior" may help resolve one of the puzzles in the literature we have just surveyed. In a widely cited natural experiment, the imposition of fines on parents arriving late to pick up their children at day care centers in Haifa resulted in a doubling of the number of tardy pickups (Gneezy and Rustichini 2000a). But the small tax on plastic grocery bags enacted in Ireland in 2002 had the opposite effect: in two weeks it resulted in a 94 percent decline in their use and appeared to crowd in social preferences (Rosenthal 2008).

The contrast is instructive. In the Haifa case, the experimenters (respecting standard experimental protocols) provided no justification for the introduction of the fine on the tardy parents. Moreover the parents' occasional lateness could have occurred for reasons beyond their control, rather than as the result of a deliberate disregard for the inconvenience it caused the teachers. Finally, lateness was not so common as to be widely broadcast to the other parents. By contrast, the introduction of the Irish plastic bag tax was preceded by a substantial publicity campaign, and the use of the bags required a deliberate choice made in a highly public condition. In the Irish case, as in the experiment by Vertova and Galbiati (2010) mentioned in section 9, the monetary incentive was introduced jointly with a message of explicit social obligation, and it apparently served as a reminder of the larger social costs of the use and disposal of the bags.

The same message comes from a voting study. In Switzerland, the removal of a negligible fine for not voting significantly reduced voting turnout; but a considerable reduction in the cost of voting (by allowing balloting by mail) had no effect on turnout.

The implication is that the fine for not voting encouraged turnout not as an incentive (by affecting the costs of not voting) but rather as a message of the importance of one's civic duty (Funk 2007).

The fact that fines often work more as messages than as incentives poses a problem for the sophisticated planner because the same intervention may bear radically different messages in different cultures. Bohnet and her coauthors implemented a Trust Game in which in one treatment the investor had the option of reducing the payoffs of trustees who betrayed their trust (Bohnet et al. 2010). Compared to the treatment in which this so-called "revenge" option was not available, when they had the revenge option a substantially larger fraction of Saudi investors trusted their partner, while a substantially smaller fraction of American investors trusted. Making trust more incentive compatible thus had diametrically opposed effects in the two cultures.

12. *Conclusion: Are Incentives to Blame?*

Is there a simple lesson for public policy? We think there is. Titmuss was right that incentives sometimes crowd out non-economic motives, and this may degrade economic performance. But Titmuss and the literature that followed him targeted incentives per se as the cause of crowding out and recommended a reduced role for incentives in the governance of economic interactions.

Both the diagnosis and the policy implication are wrong. Crowding out, as we have seen, may require greater, not lesser use of incentives. And perhaps more important: fines, subsidies, and other monetary incentives per se may not be the culprit. What accounts for crowding out, we believe, is the meaning of the fines or subsidies to the target of the incentives; and this depends on the social relationships among the actors,

the information the incentive provides, and the preexisting normative frameworks of the actors. This is the message of the contrast between the Irish grocery bag tax and the Haifa fines for tardiness, along with the fact that fines imposed on low contributors by peers in Public Goods Games have positive effects while fines imposed by principals on agents sometimes backfire. In addition, incentives chosen by agents themselves (for example, by majority rule of team members), may have a more positive effect on individual performance than if they are imposed (Mellizo, Carpenter, and Matthews 2011).

Fines deployed either to exploit or to control the target (or that give this appearance or that have this effect) are likely to be less effective than they would under separability and may even be counterproductive. The reason, we think, is that they activate the target's desire to constitute himself or herself as a dignified and autonomous individual who is treated fairly by others. It is this constitutive motive that sometimes trumps the acquisitiveness tapped by the incentive, and that leads to a contrary response. The same incentives deployed by individuals who do not stand to benefit personally, and that are intended to foster pro-social behavior are more likely to be complements rather than substitutes for social preferences, crowding them in rather than out. They do this by activating rather than diminishing the target's constitutive motives such as the desire to be treated fairly and to treat others fairly, to be a good member of a community, and the feeling of shame when others regard one as having failed in this.

Present experimental and other evidence give insufficient guidance to the planner who wishes to know ex ante, the effects of the incentives he is considering implementing. But on the basis of what we do know a good rule might be the following: The policy package of which the incentives are part should let the target understand that the desired

modification in her actions will serve to implement an outcome that is socially beneficial so that that the target is more likely to endorse the purpose of the incentive, rather than being offended by it as either unjust or a threat to her autonomy or in some other way reflecting badly on the intentions of the planner.

REFERENCES

Aaron, Henry J. 1994. "Distinguished Lecture on Economics in Government: Public Policy, Values, and Consciousness." *Journal of Economic Perspectives* 8 (2): 3–21.

Akerlof, George A. 1984. *An Economic Theorist's Book of Tales*. Cambridge and New York: Cambridge University Press.

Akerlof, George A., and Rachel E. Kranton. 2010. *Identity Economics: How Our Identities Shape Our Work, Wages, and Well-Being*. Princeton and Oxford: Princeton University Press.

Al-Ubaydli, Omar, Daniel Houser, John Nye, Maria Pia Paganelli, and Xiaofei (Sophia) Pan. 2011. "The Causal Effect of Market Participation on Trust: An Experimental Investigation Using Randomized Control." Unpublished.

Andreoni, James. 1990. "Impure Altruism and Donations to Public Goods: A Theory of Warm-Glow Giving?" *Economic Journal* 100 (401): 464–77.

Angrist, Joshua, and Victor Lavy. 2009. "The Effects of High Stakes High School Achievement Awards: Evidence from a Randomized Trial." *American Economic Review* 99 (4): 1384–1414.

Ariely, Dan, Anat Bracha, and Stephan Meier. 2009. "Doing Good or Doing Well? Image Motivation and Monetary Incentives in Behaving Prosocially." *American Economic Review* 99 (1): 544–55.

Ariely, Dan, Uri Gneezy, George Loewenstein, and Nina Mazar. 2009. "Large Stakes and Big Mistakes." *Review of Economic Studies* 76 (2): 451–69.

Arrow, Kenneth J. 1971. "Political and Economic Evaluation of Social Effects and Externalities." In *Frontiers of Quantitative Economics*, edited by Michael Intriligator, 3–23. Amsterdam: North-Holland.

Arrow, Kenneth J. 1972. "Gifts and Exchanges." *Philosophy and Public Affairs* 1 (4): 343–62.

Bandiera, Oriana, Iwan Barankay, and Imran Rasul. 2005. "Social Preferences and the Response to Incentives: Evidence from Personnel Data." *Quarterly Journal of Economics* 120 (3): 917–62.

Bandura, Albert. 1991. "Social Cognitive Theory of Moral Thought and Action." In *Handbook of Moral Behavior and Development, Volume 1: Theory*, edited by William M. Kurtines and Jacob L. Gewirtz, 45–103. Hillsdale, NJ: Erlbaum.

Baran, Nicole M., Paola Sapienza, and Luigi Zingales. 2010. "Can We Infer Social Preferences from the Lab? Evidence from the Trust Game." National Bureau of Economic Research Working Paper 15654.

Bar-Gill, Oren, and Chaim Fershtman. 2005. "Public Policy with Endogenous Preferences." *Journal of Public Economic Theory* 7 (5): 841–57.

Barr, Abigail. 2001. "Social Dilemmas and Shame-Based Sanctions: Experimental Results from Rural Zimbabwe." University of Oxford Centre for the Study of African Economies Working Paper 2001.11.

Barr, Abigail. 2004. "The Effects of Birthplace and Current Context on Other-Regarding Preferences in ACCRA." Unpublished.

Barr, Abigail, et al. 2009. "Homo Aequalis: A Cross-Society Experimental Analysis of Three Bargaining Games." University of Oxford Department of Economics Discussion Paper 422.

Becker, Gary S. 1976. "Altruism, Egoism, and Genetic Fitness: Economics and Sociobiology." *Journal of Economic Literature* 14 (3): 817–26.

Belkin, Douglas. 2002. "Boston Firefighters Sick—or Tired of Working 15-day Allowance Seen Fueling Call-Ins." *Boston Globe*, January 18, B1.

Bellemare, Charles, Sabine Kröger, and Arthur van Soest. 2008. "Measuring Inequity Aversion in a Heterogeneous Population Using Experimental Decisions and Subjective Probabilities." *Econometrica* 76 (4): 815–39.

Benabou, Roland, and Jean Tirole. 2003. "Intrinsic and Extrinsic Motivation." *Review of Economic Studies* 70 (3): 489–520.

Benabou, Roland, and Jean Tirole. 2006. "Incentives and Prosocial Behavior." *American Economic Review* 96 (5): 1652–78.

Bentham, Jeremy. 1970. *An Introduction to the Principles of Morals and Legislation*. London: Athlone Press, 1789.

Benz, Matthias, and Stephan Meier. 2008. "Do People Behave in Experiments as in the Field? Evidence from Donations." *Experimental Economics* 11 (3): 268–81.

Besley, Timothy J., Gwyn Bevan, and Konrad B. Burchardi. 2009. "Naming & Shaming: The Impacts of Different Regimes on Hospital Waiting Times in England and Wales." Center for Economic and Policy Research Discussion Paper 7306.

Bewley, Truman F. 1999. *Why Wages Don't Fall During a Recession*. Cambridge and London: Harvard University Press.

Birch, Leann L., and D. W. Marlin. 1982. "I Don't Like It; I Never Tried It: Effects of Exposure on Two-Year-Old Children's Food Preferences." *Appetite* 3 (4): 353–60.

Bisin, Alberto, and Thierry Verdier. 2001. "The Economics of Cultural Transmission and the Dynamics of Preferences." *Journal of Economic Theory* 97 (2): 298–319.

Bisin, Alberto, and Thierry Verdier. 2011. "The Economics of Cultural Transmission and Socialization." In *Handbook of Social Economics, Volume 1A*, edited by Jess Benhabib, Matthew O. Jackson, and Alberto Bisin, 339–416. Amsterdam and San Diego: Elsevier,

North-Holland.

Bliss, C. J. 1972. "Review of R. M. Titmuss 'The Gift Relationship: From Human Blood to Social Policy.'" *Journal of Public Economics* 1 (1): 162–65.

Bloom, Paul. 2010. *How Pleasure Works: The New Science of Why We Like What We Like*. New York and London: Norton.

Bohnet, Iris, and Yael Baytelman. 2007. "Institutions and Trust: Implications for Preferences, Beliefs and Behavior." *Rationality and Society* 19 (1): 99–135.

Bohnet, Iris, Bruno S. Frey, and Steffen Huck. 2001. "More Order with Less Law: On Contract Enforcement, Trust, and Crowding." *American Political Science Review* 95 (1): 131–44.

Bohnet, Iris, Fiona Greig, Benedikt Herrmann, and Richard Zeckhauser. 2008. "Betrayal Aversion: Evidence from Brazil, China, Oman, Switzerland, Turkey, and the United States." *American Economic Review* 98 (1): 294–310.

Bohnet, Iris, Benedikt Herrmann, Mohamad Al-Ississ, Andrea Robbett, Khalid Al-Yahia, and Richard Zeckhauser. 2010. "The Elasticity of Trust: How to Promote Trust in the Arab Middle East and the United States." Harvard University Kennedy School of Government Research Working Paper 10-031.

Borges, Georg, and Bernd Irlenbusch. 2007. "Fairness Crowded Out by Law: An Experimental Study on Withdrawal Rights." *Journal of Institutional and Theoretical Economics* 163 (1): 84–101.

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

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

Bowles, Samuel. 2004. *Microeconomics: Behavior, Institutions, and Evolution*. Princeton and Oxford: Princeton University Press; New York: Russell Sage Foundation.

Bowles, Samuel. 2008. "Policies Designed for Self-Interested Citizens May Undermine 'the Moral Sentiments': Evidence from Economic Experiments." *Science* 320 (5883): 1605–09.

Bowles, Samuel. 2011. "Is Liberal Society a Parasite on Tradition?" *Philosophy & Public Affairs* 39 (1): 46–81.

Bowles, Samuel, and Herbert Gintis. 2006. "Prosocial Emotions." In *The Economy as a Complex Evolving System III: Current Perspectives and Future Directions*, edited by Lawrence Blume and Steven Durlauf, 339–66. Oxford and New York: Oxford University Press.

Bowles, Samuel, and Herbert Gintis. 2011. *A Cooperative Species: Human Reciprocity and Its Evolution*. Princeton and Oxford: Princeton University Press.

Bowles, Samuel, and Sung-Ha Hwang. 2008. "Social Preferences and Public Economics: Mechanism Design When Social Preferences Depend on Incentives." *Journal of Public Economics* 92 (8–9): 1811–20.

Boyd, Robert, and Peter J. Richerson. 1985. *Culture and the Evolutionary Process*. Chicago and London: University of Chicago Press.

Burks, Stephen, Jeffrey Carpenter, and Lorenz Goette. 2009. "Performance Pay and Worker Cooperation: Evidence from an Artefactual Field Experiment." *Journal of Economic Behavior and Organization* 70 (3): 458–69.

Cardenas, Juan Camilo. 2004. "Norms from Outside and from Inside: An Experimental Analysis on the Governance of Local Ecosystems." *Forest Policy and Economics* 6 (3–4): 229–41.

Cardenas, Juan Camilo. 2005. "Groups, Commons and Regulations: Experiments with Villagers and Students in Colombia." In *Psychology, Rationality and Economic Behavior: Challenging Standard Assumptions*, edited by Bina Agarwal and Alessandro Vercelli, 242–70. Houndmills, U.K. and New York: Palgrave Macmillan.

Cardenas, Juan Camilo. 2011. "Social Norms and Behavior in the Local Commons as Seen through the Lens of Field Experiments." *Environmental and Resource Economics* 48 (3): 451–85.

Cardenas, Juan Camilo, John K. Stranlund, and Cleve Willis. 2000. "Local Environmental Control and Institutional Crowding-Out." *World Development* 28 (10): 1719–33.

Carpenter, Jeffrey, Samuel Bowles, Herbert Gintis, and Sung-Ha Hwang. 2009. "Strong Reciprocity and Team Production: Theory and Evidence." *Journal of Economic Behavior and Organization* 71 (2): 221–32.

Carpenter, Jeffrey, Stephen Burks, and Eric Verhoogen. 2005. "Comparing Students to Workers: The Effects of Social Framing on Behavior in Distribution Games." In *Field Experiments in Economics*, edited by Jeffrey Carpenter, Glenn W. Harrison, and John A. List, 261–89. Amsterdam and San Diego: Elsevier.

Carpenter, Jeffrey, Cristina Connolly, and Caitlin Knowles Myers. 2008. "Altruistic Behavior in a Representative Dictator Experiment." *Experimental Economics* 11 (3): 282–98.

Carpenter, Jeffrey, and Caitlin Knowles Myers. 2010. "Why Volunteer? Evidence on the Role of Altruism, Image, and Incentives." *Journal of Public Economics* 94 (11–12): 911–20.

Carpenter, Jeffrey, and Erika Seki. 2010. "Do Social Preferences Increase Productivity? Field Experimental Evidence from Fishermen in Toyama Bay." *Economic Inquiry* 49 (2): 612–30.

Cavalli-Sforza, Luigi, and Marcus W. Feldman. 1981. *Cultural Transmission and Evolution: A Quantitative Approach*. Princeton and Oxford: Princeton University Press.

Cervellati, Matteo, Joan Esteban, and Laurence Kranich. 2010. "Work Values, Endogenous Sentiments and Redistribution." *Journal of Public Economics* 94 (9–10): 612–27.

Cleave, Blair L., Nikos Nikiforakis, and Robert Slonim. 2010. "Is There Selection Bias in Laboratory Experiments?" University of Sydney Economics Working Paper 2010-1.

Cohn, Alain, Ernst Fehr, and Lorenz Goette. 2011. "Fair Wages and Effort: Evidence from a Field Experiment." Unpublished.

Cooley, Charles Horton. 1902. *Human Nature and the Social Order*. New York: Charles Scribner's Sons.

Cooter, Robert. 2000. "Do Good Laws Make Good Citizens? An Economic Analysis of Internalized Norms." *Virginia Law Review* 86 (8): 1577–1601.

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

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

Deci, Edward L, and Richard M. Ryan. 1985. *Intrinsic Motivation and Self-Determination in Human Behavior*. New York: Plenum Press.

DellaVigna, Stefano. 2009. "Psychology and Economics: Evidence from the Field." *Journal of Economic Literature* 47 (2): 315–72.

Dickinson, David, and Marie-Claire Villeval. 2008. "Does Monitoring Decrease Work Effort? The Complementarity between Agency and Crowding-Out Theories." *Games and Economic Behavior* 63 (1): 56–76.

Drago, Francesco, Roberto Galbiati, and Pietro Vertova. 2009. "The Deterrent Effects of Prison: Evidence from a Natural Experiment." *Journal of Political Economy* 117 (2): 257–80.

Dufwenberg, Martin, Paul Heidhues, Georg Kirchsteiger, Frank Riedel, and Joel Sobel. 2011. "Other-Regarding Preferences in General Equilibrium." *Review of Economic Studies* 78 (2): 613–39.

Ellingsen, Tore, and Magnus Johannesson. 2008. "Pride and Prejudice: The Human Side of Incentive Theory." *American Economic Review* 98 (3): 990–1008.

Ellingsen, Tore, Magnus Johannesson, Johanna Möllerström, and Sara Munkhammar. 2011. "Social Framing Effects: Preferences or Beliefs?" http://www2.hhs.se/personal/ellingsen/pdf/FramingGEB1.pdf.

Elster, Jon. 1998. "Emotions and Economic Theory." *Journal of Economic Literature* 36 (1): 47–74.

Ertan, Arhan, Talbot Page, and Louis Putterman. 2009. "Who to Punish? Individual Decisions and Majority Rule in Mitigating the Free Rider Problem." *European Economic Review* 53 (5): 495–511.

Falk, Armin, Ernst Fehr, and Urs Fischbacher. 2005. "Driving Forces Behind Informal Sanctions." *Econometrica* 73 (6): 2017–30.

Falk, Armin, Ernst Fehr, and Christian Zehnder. 2006. "Fairness Perceptions and Reservation Wages—the Behavioral Effects of Minimum Wage Laws." *Quarterly Journal of Economics* 121 (4): 1347–81.

Falk, Armin, Simon Gächter, and Judit Kovacs. 1999. "Intrinsic Motivation and Extrinsic Incentives in a Repeated Game with Incomplete Contracts." *Journal of Economic Psychology* 20 (3): 251–84.

Falk, Armin, and James J. Heckman. 2009. "Lab Experiments Are a Major Source of Knowledge in the Social Sciences." *Science* 326 (5952): 535–38.

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

Falk, Armin, Stephan Meier, and Christian Zehnder. 2011. "Did We Overestimate the Role of Social Preferences? The Case of Self-Selected Student Samples." Institute for the Study of Labor Discussion Paper 5475.

Falkinger, Josef. 1996. "Efficient Private Provision of Public Goods by Rewarding Deviations from Average." *Journal of Public Economics* 62 (3): 413–22.

Falkinger, Josef, Ernst Fehr, Simon Gächter, and Rudolf Winter-Ebmer. 2000. "A Simple Mechanism for the Efficient Provision of Public Goods: Experimental Evidence." *American Economic Review* 90 (1): 247–64.

Fehr, Ernst, and Simon Gächter. 2000. "Cooperation and Punishment in Public Goods Experiments." *American Economic Review* 90 (4): 980–94.

Fehr, Ernst, and Simon Gächter. 2002a. "Altruistic Punishment in Humans." *Nature* 415: 137–40.

Fehr, Ernst, and Simon Gächter. 2002b. "Do Incentive Contracts Undermine Voluntary Cooperation?" University of Zurich Institute for Empirical Research in Economics Working Paper 34.

Fehr, Ernst, Simon Gächter, and Georg Kirchsteiger. 1997. "Reciprocity as a Contract Enforcement Device: Experimental Evidence." *Econometrica* 65 (4): 833–60.

Fehr, Ernst, and Lorenz Goette. 2007. "Do Workers Work More If Wages Are High? Evidence from a Randomized Field Experiment." *American Economic Review* 97 (1): 298–317.

Fehr, Ernst, Alexander Klein, and Klaus M. Schmidt. 2007. "Fairness and Contract Design." *Econometrica* 75 (1): 121–54.

Fehr, Ernst, and Andreas Leibbrandt. 2011. "A Field Study on Cooperativeness and Impatience in the Tragedy of the Commons." *Journal of Public Economics* 95 (9–10): 1144–55.

Fehr, Ernst, and John A. List. 2004. "The Hidden Costs and Returns of Incentives—Trust and Trustworthiness among CEOs." *Journal of the European Economic Association* 2 (5): 743–71.

Fehr, Ernst, and Bettina Rockenbach. 2003. "Detrimental Effects of Sanctions on Human Altruism." *Nature* 422 (6928): 137–40.

Fehr, Ernst, and Klaus M. Schmidt. 1999. "A Theory of Fairness, Competition, and Cooperation." *Quarterly Journal of Economics* 114 (3): 817–68.

Fehr, Ernst, and Klaus M. Schmidt. 2007. "Adding a Stick to the Carrot? The Interaction of Bonuses and Fines." *American Economic Review* 97 (2): 177–81.

Fershtman, Chaim, and Aviad Heifetz. 2006. "Read My Lips, Watch for Leaps: Preference Equilibrium and Political Instability." *Economic Journal* 116 (508): 246–65.

Fiske, Alan P. 1992. "The Four Elementary Forms of Sociality: Framework for a Unified Theory of Social Relations." *Psychological Review* 99 (4): 689–723.

Frey, Bruno S. 1993. "Does Monitoring Increase Work Effort? The Rivalry with Trust and Loyalty." *Economic Inquiry* 31 (4): 663–70.

Frey, Bruno S. 1997. "A Constitution for Knaves Crowds Out Civic Virtues." *Economic Journal* 107 (443): 1043–53.

Frey, Bruno S., and Reto Jegen. 2001. "Motivation Crowding Theory." *Journal of Economic Surveys* 15 (5): 589–611.

Frey, Bruno S., and Felix Oberholzer-Gee. 1997. "The Cost of Price Incentives: An Empirical Analysis of Motivation Crowding-Out." *American Economic Review* 87 (4): 746–55.

Frey, Bruno S., and Alois Stutzer. 2008. "Environmental Morale and Motivation." In *The Cambridge Handbook of Psychology and Economic Behaviour*, edited by Alan Lewis, 406–28. Cambridge and New York: Cambridge University Press.

Fryer, Roland G. 2011. "Financial Incentives and Student Achievement: Evidence from Randomized Trials" *Quarterly Journal of Economics* 126 (4): 1755–98.

Funk, Patricia. 2007. "Is There An Expressive Function of Law? An Empirical Analysis of Voting Laws with Symbolic Fines." *American Law and Economics Review* 9 (1): 135–59.

Gächter, Simon, and Armin Falk. 2002. "Reputation and Reciprocity: Consequences for the Labour Relation." *Scandinavian Journal of Economics* 104 (1): 1–26.

Gächter, Simon, and Ernst Fehr. 1999. "Collective Action as a Social Exchange." *Journal of Economic Behavior and Organization* 39 (4): 341–69.

Gächter, Simon, Esther Kessler, and Manfred Königstein. 2011. "The Roles of Incentives and Voluntary Cooperation for Contractual Compliance" University of Nottingham Centre for Decision Research and Experimental Economics Discussion Paper 2011-06.

Gächter, Simon, Daniele Nosenzo, and Martin Sefton. 2010. "The Impact of Social Comparisons on Reciprocity." University of Nottingham Centre for Decision Research and Experimental Economics Discussion Paper 2010-10.

Galbiati, Roberto, and Pietro Vertova. 2008. "Obligations and Cooperative Behaviour in Public Good Games." *Games and Economic Behavior* 64 (1): 146–70.

Ginges, Jeremy, Scott Atran, Douglas Medin, and Khalil Shikaki. 2007. "Sacred Bounds on Rational Resolution of Violent Political Conflict." *Proceedings of the National Adademy of Sciences* 104 (18): 7357–60.

Gneezy, Uri. 2003. "The W Effect of Incentives." Unpublished.

Gneezy, Uri, Stephan Meier, and Pedro Rey-Biel. 2011. "When and Why Incentives (Don't) Work to Modify Behavior." *Journal of Economic Perspectives* 25 (4): 191–210.

Gneezy, Uri, and Aldo Rustichini. 2000a. "A Fine Is a Price." *Journal of Legal Studies* 29 (1): 1–17.

Gneezy, Uri, and Aldo Rustichini. 2000b. "Pay Enough or Don't Pay at All." *Quarterly Journal of Economics* 115 (3): 791–810.

Grant, Ruth W. 2011. *Strings Attached: Untangling the Ethics of Incentives*. Princeton and Oxford: Princeton University Press.

Greenberger, Scott S. 2003. "Sick Day Abuses Focus of Fire Talks." *Boston Globe*, September 17, B7.

Güth, Werner, and Hartmut Kliemt. 1994. "Competition or Co-operation: On the Evolutionary Economics of Trust, Exploitation and Moral Attitudes." *Metroeconomica* 45 (2): 155–87.

Healy, Kieran. 2006. *Last Best Gifts: Altruism and the Market for Human Blood and Organs*. Chicago and London: University of Chicago Press.

Heifetz, Aviad, Ella Segev, and Eric Talley. 2007. "Market Design with Endogenous Preferences." *Games and Economic Behavior* 58 (1): 121–53.

Helsley, Robert W., and William C. Strange. 2000. "Social Interactions and the Institutions of Local Government." *American Economic Review* 90 (5): 1477–90.

Henrich, Joseph, et al. 2005. "'Economic Man' in Cross-Cultural Perspective: Behavioral Experiments in 15 Small-Scale Societies." *Behavioral and Brain Sciences* 28 (6): 795–815.

Henrich, Joseph, et al. 2010. "Markets, Religion, Community Size, and the Evolution of Fairness and Punishment." *Science* 327 (5972): 1480–84.

Herrmann, Benedikt, and Henrik Orzen. 2008. "The Appearance of *Homo Rivalis*: Social Preferences and the Nature of Rent Seeking." University of Nottingham Centre for Decision Research and Experimental Economics Discussion Paper 2008-10.

Herrmann, Benedikt, Christian Thöni, and Simon Gächter. 2008. "Antisocial Punishment across Societies." *Science* 319 (5868): 1362–67.

Heyman, James, and Dan Ariely. 2004. "Effort for Payment: A Tale of Two Markets." *Psychological Science* 15 (11): 787–93.

Hirschman, Albert O. 1977. *The Passions and the Interests: Political Arguments for Capitalism before Its Triumph*. Princeton and Oxford: Princeton University Press.

Hirschman, Albert O. 1985. "Against Parsimony: Three Easy Ways of Complicating Some Categories of Economic Discourse." *Economics and Philosophy* 1 (1): 7–21.

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

Holmås, Tor Helge, Egil Kjerstad, Hilde Lurås, and Odd Rune Straume. 2010. "Does Monetary Punishment Crowd Out Pro-social Motivation? A Natural Experiment on Hospital Length of Stay." *Journal of Economic Behavior and Organization* 75 (2): 261–67.

Hopfensitz, Astrid, and Ernesto Reuben. 2009. "The Importance of Emotions for the Effectiveness of Social Punishment." *Economic Journal* 119 (540): 1534–59.

Houser, Daniel, Erte Xiao, Kevin McCabe, and Vernon Smith. 2008. "When Punishment Fails: Research on Sanctions, Intentions and Non-cooperation." *Games*

and Economic Behavior 62 (2): 509–32.

Hwang, Sung-Ha, and Samuel Bowles. 2011a. "Optimal Incentives with State-Dependent Preferences." Unpublished.

Hwang, Sung-Ha, and Samuel Bowles. 2011b. "The Sophisticated Planner's Dilemma: Optimal Incentives with Endogenous Preferences." Unpublished.

Irlenbusch, Bernd, and Gabriele K. Ruchala. 2008. "Relative Rewards within Team-Based Compensation." *Labour Economics* 15 (2): 141–67.

Irlenbusch, Bernd, and Dirk Sliwka. 2005. "Incentives, Decision Frames, and Motivation Crowding Out—An Experimental Investigation." Institute for the Study of Labor Discussion Paper 1758.

Karlan, Dean S. 2005. "Using Experimental Economics to Measure Social Capital and Predict Financial Decisions." *American Economic Review* 95 (5): 1688–99.

Kessler, Esther. 2008. "Behavioural Economics of Performance Incentives." University of Nottingham School of Economics.

Kosfeld, Michael, Akira Okada, and Arno Riedl. 2009. "Institution Formation in Public Goods Games." *American Economic Review* 99 (4): 1335–55.

Kreps, David M. 1997. "Intrinsic Motivation and Extrinsic Incentives." *American Economic Review* 87 (2): 359–64.

Kroll, Stephan, Todd L. Cherry, and Jason F. Shogren. 2007. "Voting, Punishment, and Public Goods." *Economic Inquiry* 45 (3): 557–70.

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

Lazear, Edward P. 2000. "Performance Pay and Productivity." *American Economic Review* 90 (5): 1346–61.

Leibbrandt, Andreas, Uri Gneezy, and John A. List. 2010. "Ode to the Sea: The Socio-ecological Underpinnings of Social Norms." Unpublished.

Leider, Stephen, Markus M. Möbius, Tanya Rosenblat, and Quoc-Anh Do. 2009. "Directed Altruism and Enforced Reciprocity in Social Networks." *Quarterly Journal of Economics* 124 (4): 1815–51.

Lepper, Mark R., Gerald Sagotsky, Janet L. Dafoe, and David Greene. 1982. "Consequences of Superfluous Social Constraints: Effects on Young Children's Social Inferences and Subsequent Intrinsic Interest." *Journal of Personality and Social Psychology* 42 (1): 51–65.

Levine, David K. 1998. "Modeling Altruism and Spitefulness in Experiments." *Review of Economic Dynamics* 1 (3): 593–622.

Levitt, Steven D., and John A. List. 2007. "What Do Laboratory Experiments Measuring Social Preferences Reveal about the Real World?" *Journal of Economic Perspectives* 21 (2): 153–74.

Li, Jian, Erte Xiao, Daniel Houser, and P. Read Montague. 2009. "Neural Responses to Sanction Threats in Two-Party Economic Exchange." *Proceedings of the National Academy of Sciences* 106 (39): 16835–40.

List, John A. 2004. "Young, Selfish and Male: Field Evidence of Social Preferences." *Economic Journal* 114 (492): 121–49.

Loewenstein, George. 2000. "Emotions in Economic Theory and Economic Behavior." *American Economic Review* 90 (2): 426–32.

Lopez, Maria Claudia, James J. Murphy, John M. Spraggon, and John K. Stranlund. 2012. "Comparing the Effectiveness of Regulation and Pro-social Emotions to Enhance Cooperation: Experimental Evidence from Fishing Communities in Colombia." *Economic Inquiry* 50 (1): 131–42.

Lucas, Robert E. 1976. "Econometric Policy Evaluation: A Critique." *Carnegie–Rochester Conference Series on Public Policy* 1: 19–46.

MacLeod, W. Bentley. 2007. "Can Contract Theory Explain Social Preferences?" *American Economic Review* 97 (2): 187–92.

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

Meier, Stephan. 2007a. "Do Subsidies Increase Charitable Giving in the Long Run? Matching Donations in a Field Experiment." *Journal of the European Economic Association* 5 (6): 1203–22.

Meier, Stephan. 2007b. "A Survey of Economic Theories and Field Evidence on Pro-social Behavior." In *Economics and Psychology: A Promising New Cross-Disciplinary Field*, edited by Bruno S. Frey and Alois Stutzer, 51–88. Cambridge and London: MIT Press.

Mellizo, Philip, Jeffrey Carpenter, and Peter Hans Matthews. 2011. "Workplace Democracy in the Lab." Unpublished.

Mellstrom, Carl, and Magnus Johannesson. 2008. "Crowding Out in Blood Donation: Was Titmuss Right?" *Journal of the European Economic Association* 6 (4): 845–63.

Mill, John Stuart. 1867. *Principles of Political Economy with Some of Their Applications to Social Philosophy*. New York; Bombay and Culcutta: Longmans, Green and Co., 1848.

Mulder, Laetitia B., Eric van Dijk, David De Cremer, and Henk A. M. Wilke. 2006. "Undermining Trust and Cooperation: The Paradox of Sanctioning Systems in Social Dilemmas." *Journal of Experimental Social Psychology* 42 (2): 147–62.

Murphy, Sheila T., Jennifer L. Monahan, and Robert B. Zajonc. 1995. "Additivity of Nonconscious Affect: Combined Effects of Priming and Exposure." *Journal of Personality and Social Psychology* 69 (4): 589–602.

Murphy, Sheila T., and Robert B. Zajonc. 1993. "Affect, Cognition, and Awareness: Affective Priming with Optimal and Suboptimal Stimulus Exposures." *Journal of Personality and Social Psychology* 64 (5): 723–39.

Rabin, Matthew. 1993. "Incorporating Fairness into Game Theory and Economics." *American Economic Review* 83 (5): 1281–1302.

Rabin, Matthew. 1998. "Psychology and Economics." *Journal of Economic Literature* 36 (1): 11–46.

Reeson, Andrew F., and John G. Tisdell. 2008. "Institutions, Motivations and Public Goods: An Experimental Test of Motivational Crowding." *Journal of Economic Behavior and Organization* 68 (1): 273–81.

Rigdon, Mary. 2009. "Trust and Reciprocity in Incentive Contracting." *Journal of Economic Behavior and Organization* 70 (1–2): 93–105.

Rodriguez-Sickert, Carlos, Ricardo Andrés Guzmán, and Juan Camilo Cardenas. 2008. "Institutions Influence Preferences: Evidence from a Common Pool Resource Experiment." *Journal of Economic Behavior and Organization* 67 (1): 215–27.

Rosenthal, Elisabeth. 2008. "Motivated by a Tax, Irish Spurn Plastic Bags." *New York Times*, February 2.

Ross, Lee, and Richard E. Nisbett. 1991. *The Person and the Situation: Perspectives of Social Psychology*. Philadelphia: Temple University Press.

Rustagi, Devesh, Stefanie Engel, and Michael Kosfeld. 2010. "Conditional Cooperation and Costly Monitoring Explain Success in Forest Commons Management." *Science* 330 (6006): 961–965.

Rutstrom, E. Elisabet. 2002. "Sparing the Rod Does Not Spoil the Child: An Experimental Study of Incentive Effects." Unpublished.

Schnedler, Wendelin, and Radovan Vadovic. 2011. "Legitimacy of Control." *Journal of Economics and Management Strategy* 20 (4): 985–1009.

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

Seabright, Paul B. 2009. "Continuous Preferences and Discontinuous Choices: How Altruists Respond to Incentives." *B.E. Journal of Theoretical Economics* 9 (1).

Serra, Danila. 2008. "Combining Top–Down and Bottom–Up Accountability: Evidence from a Bribery Experiment." University of Oxford Centre for the Study of African Economies Working Paper 2008-25.

Shinada, Mizuho, and Toshio Yamagishi. 2007. "Punishing Free Riders: Direct and Indirect Promotion of Cooperation." *Evolution and Human Behavior* 28 (5): 330–39.

Shu, Lisa L., Francesca Gino, and Max H. Bazerman. 2011. "Dishonest Deed, Clear Conscience: When Cheating Leads to Moral Disengagement and Motivated Forgetting." *Personality and Social Psychology Bulletin* 37 (3): 330–49.

Sliwka, Dirk. 2007. "Trust as a Signal of a Social Norm and the Hidden Costs of Incentive Schemes." *American Economic Review* 97 (3): 999–1012.

Sloof, Randolph, and Joep Sonnemans. 2001. "The Interaction between Explicit and Relational Incentives: An Experiment." *Games and Economic Behavior* 73 (2): 573–94.

Sobel, Joel. 2002. "Can We Trust Social Capital?" *Journal of Economic Literature* 40 (1): 139–54.

Sobel, Joel. 2005. "Interdependent Preferences and Reciprocity." *Journal of Economic Literature* 43 (2): 392–436.

Sobel, Joel. 2009. "Generous Actors, Selfish Actions: Markets with Other-Regarding Preferences." *International Review of Economics* 56 (1): 3–16.

Sobel, Joel. 2010. "Markets and Other-Regarding Preferences." http://federation.ens.fr/ydepot/semin/texte0910/SOB2010MAR.pdf.

Solow, Robert M. 1971. "Blood and Thunder: Review of the Gift Relationship: From Human Blood to Social Policy by Richard M. Titmuss." *Yale Law Journal* 80 (8): 1696–1711.

Stanca, Luca, Luigino Bruni, and Luca Corazzini. 2009. "Testing Theories of Reciprocity: Do Motivations Matter?" *Journal of Economic Behavior and Organization* 71 (2): 233–45.

Stiglitz, Joseph E. 1987. "The Causes and Consequences of the Dependence of Quality on Price." *Journal of Economic Literature* 25 (1): 1–48.

Sunstein, Cass R. 1996. "On the Expressive Function of Law." *University of Pennsylvania Law Review* 144 (5): 2021–53.

Sutter, Matthias, Stefan Haigner, and Martin G. Kocher. 2011. "Choosing the Carrot or the Stick? Endogenous Institutional Choice in Social Dilemma Situations." *Review of Economic Studies* 77 (4): 1540–66.

Taylor, Michael. 1987. *The Possibility of Cooperation*. Cambridge and New York: Cambridge University Press.

Tenbrunsel, Ann F., and David M. Messick. 1999. "Sanctioning Systems, Decision Frames, and Cooperation." *Administrative Science Quarterly* 44: 684–707.

Tirole, Jean. 1999. "Incomplete Contracts: Where Do We Stand?" *Econometrica* 67 (4): 741–81.

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

Tversky, Amos, and Daniel Kahneman. 1981. "The Framing of Decisions and the Psychology of Choice." *Science* 211 (4481): 453–58.

Tyran, Jean-Robert, and Lars P. Feld. 2006. "Achieving Compliance When Legal Sanctions Are Nondeterrent." *Scandinavian Journal of Economics* 108 (1): 135–56.

Vertova, Pietro, and Roberto Galbiati. 2010. "How Laws Affect Behavior: Obligations, Incentives and Cooperative Behavior. http://ssrn.com/paper=1615349.

Warneken, Felix, and Michael Tomasello. 2008. "Extrinsic Rewards Undermine Altruistic Tendencies in 20-month-olds." *Developmental Psychology* 44 (6): 1785–88.

Xiao, Erte, and Daniel Houser. 2011. "Punish in Public." *Journal of Public Economics* 95 (7–8): 1006–17.

Yeung, King-To, and John Levi Martin. 2003. "The Looking Glass Self: An Empirical Test and Elaboration." *Social Forces* 81 (3): 843–79.

Young, H. Peyton, and Mary A. Burke. 2001. "Competition and Custom in Economic Contracts: A Case Study of Illinois Agriculture." *American Economic Review* 91 (3): 559–73.

Zajonc, Robert B. 1968. "Attitudinal Effects of Mere Exposure." *Journal of Personality and Social Psychology* 9 (2 Part 2): 1–27.

Zhong, Chen-Bo, Vanessa K. Bohns, and Francesca Gino. 2010. "Good Lamps Are the Best Police: Darkness Increases Dishonestly and Self-Interested Behavior." *Pschological Science* 21 (3): 311–14.

Zhong, Chen-Bo, Jeffrey Loewenstein, and J. Keith Murnighan. 2007. "Speaking the Same Language: The Cooperative Effects of Labeling in the Prisoner's Dilemma." *Journal of Conflict Resolution* 51 (3): 431–56.

This article has been cited by:

1. Michael Kosfeld, Devesh Rustagi. 2015. Leader Punishment and Cooperation in Groups: Experimental Field Evidence from Commons Management in Ethiopia. *American Economic Review* **105**:2, 747-783. [Abstract] [View PDF article] [PDF with links]

2. Rajna Gibson,, Carmen Tanner,, Alexander F. Wagner. 2013. Preferences for Truthfulness: Heterogeneity among and within Individuals. *American Economic Review* **103**:1, 532-548. [Abstract] [View PDF article] [PDF with links]