THE ECONOMIC JOURNAL

The Economic Journal, **127** (*June*), 873–895. Doi: 10.1111/ecoj.12321 © 2015 Royal Economic Society. Published by John Wiley & Sons, 9600 Garsington Road, Oxford OX4 2DQ, UK and 350 Main Street, Malden, MA 02148, USA.

ASSIGNMENT PROCEDURE BIASES IN RANDOMISED POLICY EXPERIMENTS*

Gani Aldashev, Georg Kirchsteiger and Alexander Sebald

Randomised controlled trials (RCT) have gained ground as the dominant tool for studying policy interventions in many fields of applied economics. We analyse theoretically encouragement and resentful demoralisation in RCTs and show that these might be rooted in the same behavioural trait – people's propensity to act reciprocally. When people are motivated by reciprocity, the choice of assignment procedure influences the RCTs' findings. We show that even credible and explicit randomisation procedures do not guarantee an unbiased prediction of the impact of policy interventions; however, they minimise any bias relative to other less transparent assignment procedures.

Randomised controlled trials (hereafter RCTs) have gained ground as the dominant tool for studying the effects of policy interventions on outcomes of interest in many fields of applied economics, most notably in labour economics, development economics and public finance. Researchers have used RCTs to study such diverse questions as the effects of conditional cash transfers to poor families on education and on health of children in Mexico (Gertler, 2004; Schultz, 2004), of vouchers for private schooling on school completion rates in Colombia (Angrist, *et al.* 2002, 2006), of publicly released audits on electoral outcomes in Brazil (Ferraz and Finan, 2008), of incremental cash investments on the profitability of small enterprises in Sri Lanka (De Mel *et al.*, 2008), of income subsidies on work incentives in Canada (Card and Hyslop, 2005; Card and Robins, 2005; Michalopoulos *et al.*, 2005), of saving incentives on the saving decisions of low and middle-income families in the US (Duflo *et al.*, 2006), and of the introduction of microfinance institutions on small business start-ups and consumption patterns in India (Banerjee *et al.*, 2010).

Typically, RCTs are used for *ex ante* programme evaluation purposes. To evaluate *ex ante* the effect of a general introduction of a policy or development NGO intervention on some social or economic outcome, researchers assign individuals (or other units under study, e.g. schools or villages) into a treatment and a control group. The individuals in the treatment group receive the policy 'treatment' and subsequently their behaviour is compared to that of the individuals in the control group. The observed difference between the outcomes in the treatment and the control group is used as a predictor for the effect of a general introduction of the

* Corresponding author: Alexander Sebald, Department of Economics, University of Copenhagen, Øster Farimagsgade 5, 1353 Copenhagen, Denmark. Email: alexander.sebald@econ.ku.dk.

We thank the editor of the Economic Journal and two anonymous referees whose comments and suggestions improved the article to a great extent. We also thank Angus Deaton, Pramila Krishnan, Claudia Senik and seminar participants at Cambridge, Konstanz, Helsinki and the Paris School of Economics for very useful comments. Kirchsteiger acknowledges financial support from the FRFC project on 'Preference dynamics in adaptive networks' (project n 2.4614.12).

programme. Based on the experimental results, the programme might be generally adopted or not.[1]

Notwithstanding the empirical importance of RCTs in evaluating the impact of policy interventions, there also exists an old debate concerning factors that might mitigate or compromise their internal and external validity. Factors that have been shown to potentially influence the validity of RCTs are among others the randomisation bias (Heckman, 1991), the Hawthorne and John Henry effect – Levitt and List (2011) and Duflo *et al.* (2008), and the placebo effect (Malani 2006).

In our analysis, we concentrate on the Hawthorne and John Henry effect. Interestingly, although the start of the debate about these two effects dates back to the 1950s and 1970s respectively, one of the difficulties in analysing their character and importance is the absence of a formal definition.[2] A broad verbal definition of the Hawthorne and John Henry effect is provided by Duflo *et al.* (2008, p. 3951):

> Changes in behaviour among the treatment group are called Hawthorne effects, while changes in behaviour among the comparison group are called John Henry effects. The treatment group may be grateful to receive a treatment and conscious of being observed, which may induce them to alter their behaviour for the duration of the experiment (e.g. working harder to make it a success). The comparison group may feel offended to be a comparison group and react by also altering their behaviour (for example, teachers in the comparison group for an evaluation may 'compete' with the treatment teachers or, on the contrary, decide to slack off).

In line with the above definition, a change in the behaviour of the control group is also well-known in psychology under the heading 'resentful demoralisation'. The phenomenon was first described in detail by Cook and Campbell in their seminal book on experimental methods (Cook and Campbell, 1979), where they list resentful demoralisation among potential threats to the internal validity of experiments in social sciences (Fetterman, 1982; Ongena, 2009):

> When an experiment is obtrusive, the reaction of a no-treatment control group or groups receiving less desirable treatments can be associated with resentment and demoralisation ... In an industrial setting the persons experiencing the less desirable treatments might retaliate by lowering productivity and company profits, while in an educational setting, teachers or students could 'lose heart' or become angry and 'act up'. Any of these forces could lead to a post-test difference between treatment and no-treatment groups, and it would be quite wrong to attribute the difference to the planned treatment ... Rather, it would be from the inadvertent resentful demoralisation experienced by the non-treatment controls' (Cook and Campbell, 1979, p. 55).

[1] See Duflo (2004) for a description of the RCT and the subsequent general implementation of PROGRESA conditional cash transfer programme for school attendance in Mexico.

[2] Different (verbal) definitions exist. To this effect Levitt and List (2011, p. 227) write: 'The Merriam-Webster dictionary offers a very different definition for a Hawthorne effect than the one cited in the Oxford English Dictionary (OED): the stimulation to output or accomplishment that results from the mere fact of being under observation'.

© 2015 Royal Economic Society.

Building on this intuition and on the recent literature on belief-dependent preferences (Geanakoplos *et al.*, 1989; Battigalli and Dufwenberg, 2009), we theoretically analyse such feelings of encouragement (of the treatment group) and resentful demoralisation (of the control group) as described by Duflo *et al.* (2008) and Cook and Campbell (1979). We show that these might be two sides of the same behavioural trait, namely people's propensity to act reciprocally. To analyse theoretically the impact of encouragement and resentful demoralisation on the validity of results generated by RCTs, we construct a simple game-theoretic model of RCTs in which agents are motivated by belief-dependent preferences. We adopt the framework suggested by Sebald (2010)[3] in which agents are willing to react positively to a particularly good treatment and negatively to a particularly bad one.[4]

Our formal analysis not only provides a clear theoretical basis that can be used to analyse feelings of encouragement and resentful demoralisation in RCTs but it also delivers intriguing insights regarding their potential character and importance. In particular, we show that it might not only be the fact that people are 'under scrutiny' or 'under observation' in RCTs that drives these biases but that the assignment procedure used to allocate people into control and treatment group crucially determines the size of these biases.

In line with Duflo *et al.*'s (2008) definition of the Hawthorne and John Henry effect we find that a reciprocal subject that is not assigned to the treatment group (while other similar agents are) feels discouraged and provides less effort than he would without the presence of a treatment group. Hence, control group subjects are particularly demotivated. On the other hand, if a participant is assigned to the treatment group (while some other subjects are not), he feels particularly encouraged to provide more effort than without the existence of the control group. Consequently, the observed difference between the outcomes of the treatment and the control groups delivers a biased prediction of the effect of a general introduction of the treatment.

The size of the bias depends crucially on the assignment procedure itself. If a subject is assigned to the control (treatment) group through a non-transparent (private) randomisation procedure, the amount of resentful demoralisation (encouragement) is particularly high. The estimate of the effect of a general introduction of the treatment under this type of randomisation procedure is unambiguously biased upwards. On the other hand, if the experimenter uses an explicit and credible randomisation mechanism, the impact of demoralisation and encouragement is lower. Hence, the problem of the upward bias in the estimate is reduced. However, an unbiased and transparent randomisation procedure might even lead to a negative bias, i.e. an underestimation of the true estimate. Our analysis reveals that no assignment procedure necessarily guarantees that the observed difference in outcomes of the control and treatment groups coincides with the true effect of a general introduction of the treatment. But an unbiased and credible randomisation procedure for allocating the

[3] Sebald (2010) generalises the reciprocity model of Dufwenberg and Kirchsteiger (2004) to settings in which moves of chance are possible. Given that randomisation in policy experiments crucially involves moves of chance, this model fits our setting particularly well.

[4] There exists a lot of experimental evidence for this type of behaviour. For an overview, see Sobel (2005).

© 2015 Royal Economic Society.

subjects into the treatment and the control group leads to a smaller bias in the estimation of the treatment effect than less transparent assignment mechanisms.

This article contributes to the small but growing economic literature theoretically analysing the behaviour of subjects in RCTs.[5] The papers closest to ours are Philipson and Hedges (1998), Malani (2006) and Chassang *et al.* (2012).

Philipson and Hedges (1998) study a model of attrition built on the premise that treatment-group subjects in RCTs face stronger incentives to avoid type I and II errors than the researcher. Thus, they rationally decide on staying in or quitting the experiment and reveal, through attrition, their learning about (and the utility derived from) the effect of the treatment. One implication is that information about treatment preference can be inferred from the standard data on attrition rates of RCTs.

Malani (2006) builds a simple model of the placebo effect in (medical) RCTs, i.e. the effect arising purely from the subjects' response to treatment depending positively on their expectations about the value of the treatment. In his model the individual outcome is influenced both by the treatment directly and by the belief of the individual about the effectiveness of the treatment. More optimistic patients respond more strongly to treatment than the less optimistic ones. The obtained empirical estimates of the effectiveness of the treatment will be imprecise, because of the combination of the genuine treatment effect and the placebo effect. As a solution, the article proposes an experimental design with two (or more) treatment groups plus a control group, and varying the probability of obtaining the treatment across the treatment groups. Higher observed outcomes for non-treated subjects in the treatment group(s) with higher *ex ante* probability of obtaining the treatment indicate the presence of the placebo effect.

Chassang *et al.* (2012) study a related problem of identifying the effect of the treatment in a setting where there is an underlying (unobservable) heterogeneity of subjects' expectations about the effectiveness (or 'returns') of the treatment. The overall outcome depends on the actual effectiveness of the treatment but also on some (costly) effort. Since subjects' heterogenous expectations about the return effect their effort levels, the estimate of treatment's effectiveness obtained from such an experiment would be imprecise. The proposed solution relies on the mechanism-design approach and consists in letting subjects reveal their preferences over their treatment by probabilistically selecting themselves in (or out) of groups at a cost.

Our contribution differs from the above studies in that the focus of our study is the demoralisation and encouragement effect created by the assignment procedure, i.e. the Hawthorne and John Henry effect, an issue not analysed in the above literature.

In the next Section we present a simple model of policy experiments that takes demoralisation and encouragement into account. Section 2 derives the biases connected to the different randomised assignment procedures formally. In Section 3 we discuss prominent examples of policy experiments in which the choice of assignment procedure biased the measured treatment effect and describe the implication of our results for the design of RCTs. Section 4 concludes.

[5] Of course, there is a large methodological literature in empirical economics that discusses various biases that might arise in inferring the effects of a programme from observed outcomes in experimental settings. Excellent reviews are provided by Heckman and Vytlacil (2007*a*, *b*), Abbring and Heckman (2007) and Imbens and Wooldridge (2009).

© 2015 Royal Economic Society.

1. A Simple Model of Policy Experiments

Consider a policy experiment that entails giving some benefits to subjects in the treatment group. These benefits (e.g. a tool, school supplies, or job market training) constitute an input into the production function of the outcome of interest for the experimenter (e.g. agricultural productivity, learning outcomes, or likelihood of finding a job). We denote by N the size of the overall population and by n the number of those agents who are subject to the treatment. $q = n/N \in [0, 1]$ denotes the fraction of agents in the treatment group.

To concentrate on the impact of the randomised assignment procedures, we abstract from any idiosyncratic differences between the agents. Thus, all agents are identical except for their treatment status. For simplicity, we assume that the experimenter can choose between two procedures to assign individuals into the treatment and the control group:

(*i*) the experimenter can choose the n treatment-group subjects directly. This also models a closed-doors random assignment procedure, when the agents do not believe in the randomness of the assignment; and
(*ii*) the experimenter can choose an explicit randomisation procedure observable to the agents, such that each agent has the same probability q of receiving the treatment.

Since we are interested in the impact of the assignment procedure, we will not analyse the experimenter's equilibrium choice as if he were a player. Rather, we will compare the reaction of the agents to the two assignment procedures.

Formally, any subset of the overall population with n agents is a feasible action of the experimenter. The set of feasible procedures is given by all degenerate probability distributions that choose an action for sure (i.e. direct appointment of the n treatment agents) and by the procedure where the experimenter chooses the n treatment agents with the help of a public and fair lottery. Note that since all agents are equal, all these 'degenerate' procedures where the treatment agents are picked directly induce the same choices of the 'treated' as well as of the 'untreated' agents. Therefore, we restrict the analysis to a typical element of this class of procedures, denoted by d. Denoting the public randomisation procedure by r, the experimenter's set of assignment procedures is given by $P = \{d, r\}$ with p denoting a typical element of this set. Upon assignment, the chosen agents receive the treatment, whereas the other individuals do not receive it. Next, all agents choose simultaneously an effort level $e \in [0, 1]$. We assume that each agent's strategy set consists only of pure strategies, i.e. he cannot randomise over different feasible effort levels (a pure strategy equilibrium always exists, see Proposition 1 below). This restriction is common knowledge.

In most RCTs, the outcome of interest for the experimenter depends not only on the treatment itself but also on the effort level of the agents. Thus, as in Chassang *et al.* (2012), we model the outcome as depending on treatment and effort. Let the marginal success of effort be constant and denoted by t. For analytical simplicity, we assume that $t = 1$ for agents that receive the treatment and $t = 1/2$ for the other agents. Thus, the treatment makes it easier for participants to be successful. We use the variable $t \in \{1/2, 1\}$ to characterise also whether an agent is in the control group ($t = 1/2$) or

© 2015 Royal Economic Society.

in the treatment group ($t = 1$). We denote with (t, p) the *type* of the agent who is put into group t by the assignment procedure p. We restrict our attention to symmetric equilibria where all agents of the same type (t, p) choose the same effort level $e(t, p)$. Together with (the lack of) the treatment, this effort determines the success of an agent with respect to, for example, finding a job or stopping drug consumption. Formally, the success of a (t, p)-agent is given by:

$$s = t \times e(t, p). \tag{1}$$

As already mentioned, we do not analyse the experimenter's equilibrium choice as if he were a player. However, to determine the reaction of the agents to the assignment procedure, we have to specify the goal of the experimenter as perceived by the agents. In almost every policy experiment, the subjects do not know that the goal of the researcher is to evaluate the effectiveness of the policy intervention by comparing the outcomes of the treatment and control groups. If the agents knew that the effectiveness of the programme were tested and that the experimental results determine the long-run feasibility and shape of the programme, the agents' long-term strategic interests would jeopardise the validity of the experimental results. To give the randomised experiments the best shot, we abstract from such effects by assuming that the agents, unaware of the experimental character of the programme, consider the overall success, denoted by π_x, as the goal of the experimenter.[6] It depends on the effort levels chosen by the agents (which, in turn, depends on the assignment procedure), and on the group sizes:

$$\pi_x = n \times e(1, p) + (N - n) \times \frac{1}{2} \times e\left(\frac{1}{2}, p\right). \tag{2}$$

We assume that the agents are motivated by their individual success: the unemployed want to find a job, drug users want to get clean etc. Furthermore, each agent has to bear the cost of effort, which we assume to be quadratic. Disregarding the psychological payoff, a (t, p)-agent's direct (or 'material') payoff is:[7]

$$\pi[t, e(t, p)] = t \times e(t, p) - e(t, p)^2. \tag{3}$$

Both the experimenter's payoff as perceived by the agents, π_x, as well as the agents' payoff π refer to the material success of the programme. However, as we argue above, agents do not only care about their material payoffs but also about the way they are

[6] The exact form of the experimenter's goal as perceived by the agents is not important for our results. Any goal function would lead to allocation biases, as long as each agent believes that the experimenter cares about her success at least to some strictly positive extent. In this case, any effort increase is an increase in the kindness provided by the agent to the experimenter and this justifies to assume (7).

[7] As can be seen from this material payoff function, we assume that the effort cost function is the same for treated and untreated subjects. Subjects only differ in terms of productivity of effort. Alternatively, we could have measured effort such that it measures the marginal product directly. Denote by $\rho(t, p)$ the marginal product of a (t, p)-agent. This implies that $\rho(1/2, p) = 1/2e(1/2, p)$, $\rho(1, p) = e(1, p)$, and

$$\pi[t, \rho(t, p)] = \rho(t, p) - \left[\frac{1}{t}\rho(t, p)\right]^2.$$

With this effort measure, the treated and untreated agents have the same constant marginal product of effort, but they differ in the effort cost function. Obviously, all results derived below would remain unchanged by measuring the effort in this way.

© 2015 Royal Economic Society.

treated. If an agent feels treated badly, he resents the experimenter, feels discouraged, and hence, is less willing to provide effort. On the other hand, if the agent feels treated particularly well, he might feel encouraged, may want the programme to be a success, and hence provides higher effort. In other words, agents are not only concerned about their material payoff but also act reciprocally.

Crucially, whether an agent feels treated kindly or unkindly depends on how much material payoff he 'thinks' that the experimenter 'intends' to give him relative to a 'neutral' material payoff.

To model such concerns, we need to introduce first-order and second-order beliefs into the utility functions. For any t, t' and p, p', denote by $\bar{e}^{t,p}(t', p')$ the first-order belief of a (t, p)-agent about the effort choice of a (t', p')-agent. $\bar{e}^{t,p}(t, p)$ is the belief of a (t, p)-agent about the effort choice of the other agents of his own type. The first-order beliefs of a (t, p)-agent are thus summarised by:

$$\bar{e}^{t,p} = \left[\bar{e}^{t,p}(1, d), \bar{e}^{t,p}\left(\frac{1}{2}, d\right), \bar{e}^{t,p}(1, r), \bar{e}^{t,p}\left(\frac{1}{2}, r\right)\right].$$

Following Dufwenberg and Kirchsteiger (2004), we assume that every agent holds a point belief, i.e. he thinks that he knows the effort choices of the other agents for sure. This assumption, together with the pure strategy choice of all agents, implies that the set of possible first-order beliefs of a (t, p)-agent about the effort choice of a (t', p')-agent is equal to the set of pure strategies of the (t', p')-agent, i.e. $\bar{e}^{t,p} \in [0, 1]^4$.

Furthermore, we also need the second-order belief of a (t, p)-agent about the experimenter's belief concerning the effort choice of a (t', p')-agent. Denote this belief by $\bar{\bar{e}}^{t,p}(t', p')$. The second-order beliefs of a (t, p)-agent are summarised by:

$$\bar{\bar{e}}^{t,p} = \left[\bar{\bar{e}}^{t,p}(1, d), \bar{\bar{e}}^{t,p}\left(\frac{1}{2}, d\right), \bar{\bar{e}}^{t,p}(1, r), \bar{\bar{e}}^{t,p}\left(\frac{1}{2}, r\right)\right].$$

Following Dufwenberg and Kirchsteiger (2004) we again assume point beliefs. Hence, $\bar{\bar{e}}^{t,p} \in [0, 1]^4$.

Denote by $\pi_x[e(t, p), \bar{e}^{t,p}]$ the level of overall outcome or 'success' of the programme that a (t, p)-agent intends for the programme if he chooses $e(t, p)$ and he believes that the others choose $\bar{e}^{t,p}$. It is given by:

$$\pi_x[e(t,p), \bar{e}^{t,p}] = \begin{cases} e(1,p) + (n-1) \times \bar{e}^{1,p}(1,p) + (N-n) \times \dfrac{1}{2} \times \bar{e}^{1,p}(\frac{1}{2}, p) & \text{if } t = 1 \\ \dfrac{1}{2} \times e(\dfrac{1}{2}, p) + n \times \bar{e}^{\frac{1}{2},p}(1,p) + (N-n-1) \times \dfrac{1}{2} \times \bar{e}^{\frac{1}{2},p}(\dfrac{1}{2}, p) & \text{if } t = \dfrac{1}{2}. \end{cases} \tag{4}$$

Note that $\pi_x[e(t, p), \bar{e}^{t,p}]$ does not depend on the actual effort of the other agents but on the agents' belief about the other agents' effort. Any change of $e(t, p)$ does not change what the particular (t, p)-agent thinks the other agents will contribute to the overall success. This is reflected by:

$$\frac{\partial \pi_x[e(t,p), \bar{e}^{t,p}]}{\partial e(t,p)} = t.$$

© 2015 Royal Economic Society.

$\pi(\bar{\bar{e}}^{t,p})$ denotes the belief of a (t, p)-agent about the expected material payoff the experimenter intends to give him. Crucially, we assume that the agents do not hold the experimenter responsible for the outcome of the public random assignment mechanism.[8] Hence, $\pi(\bar{\bar{e}}^{t,p})$ is given by:

$$\pi(\bar{\bar{e}}^{t,p}) = \begin{cases} q \times \left[\bar{\bar{e}}^{t,r}(1,r) - \bar{\bar{e}}^{t,r}(1,r)^2\right] + (1-q) \times \left[\frac{1}{2} \times \bar{\bar{e}}^{t,r}\left(\frac{1}{2}, r\right) - \bar{\bar{e}}^{t,r}\left(\frac{1}{2}, r\right)^2\right] & \text{if } p = r \\ t \times \bar{\bar{e}}^{t,d}(t,d) - \bar{\bar{e}}^{t,d}(t,d)^2 & \text{if } p = d. \end{cases} \quad (5)$$

Note that $\pi(\bar{\bar{e}}^{1,r}) = \pi(\bar{\bar{e}}^{\frac{1}{2},r})$ whenever $\bar{\bar{e}}^{1,r} = \bar{\bar{e}}^{\frac{1}{2},r}$. In other words, when the public randomisation procedure is used and the agent's second-order beliefs are independent of his group t, the agent's beliefs about the payoff that the experimenter intends to give him are not influenced by the agent's treatment status. Furthermore, $\pi(\bar{\bar{e}}^{t,p}) \in [-\frac{1}{2}, \frac{1}{4}]$ since $e \in [0, 1]$.

We also have to specify the 'neutral' payoff $\hat{\pi}$ at which the agent regards the principal's choice of assignment procedure as being materially neutral, i.e. neither favouring nor discriminating against the agent.[9] As will be clear from the specification of the utility function below, whenever the agent thinks that the experimenter intends to give him $\hat{\pi}$, he is neither discouraged nor encouraged, and hence he simply maximises his material payoff.

Note that the expected material payoff of an agent is maximised when he is directly assigned to the treatment group. It is minimised when the agent is directly assigned to the control group. Therefore, we assume that $\hat{\pi}$ is a weighted average between the payoff that the agent thinks that the experimenter intends to give to someone directly assigned into the treatment group and the intended material payoff for an agent directly assigned into the control group. The weights are denoted by λ and $1 - \lambda$, respectively, with $\lambda \in [0,1]$:

$$\hat{\pi}(\bar{\bar{e}}^{t,p}) = \lambda \times \left[\bar{\bar{e}}^{t,p}(1,d) - \bar{\bar{e}}^{t,p}(1,d)^2\right] + (1-\lambda) \times \left[\frac{1}{2} \times \bar{\bar{e}}^{t,p}\left(\frac{1}{2}, d\right) - \bar{\bar{e}}^{t,p}\left(\frac{1}{2}, d\right)^2\right], \quad (6)$$

with $\hat{\pi}(\bar{\bar{e}}^{t,p}) \in [-\frac{1}{2}, \frac{1}{4}]$ since $e \in [0, 1]$.

The weight λ depends on the fraction of agents that are subject to the treatment, i.e. q. Whenever a randomised control trial is conducted, i.e. if $q \in (0, 1)$, the agents take the existence of both groups into account, i.e. $\lambda \in (0, 1)$. In the extreme cases when nobody (everybody) is subject to the treatment, i.e. when $q = 0$ ($q = 1$), the agents are aware of it, i.e. $\lambda = 0$ ($\lambda = 1$). Moreover, for $q \in (0, 1)$ it seems natural to assume that $\lambda = q$. However, it is well-known that people's perception about what they deserve is often self-serving. For instance, most people regard themselves as being more talented than the average (the so-called 'Lake Wobegon effect' – see Hoorens, 1993). Therefore, many individuals in the policy programme might think that they deserve

[8] This assumption gives the RCTs 'the best chance'. If this assumption fails, the publicly randomised assignment procedure would induce a level of demoralisation and encouragement similar to those under the direct assignment. As a consequence, the public randomisation procedure would induce the same kind of bias as the private randomisation.

[9] $\hat{\pi}$ plays a role similar to the 'equitable' payoffs in Rabin (1993) and Dufwenberg and Kirchsteiger (2004).

© 2015 Royal Economic Society.

the treatment more than the others, implying that $\lambda > q$. On the other hand, we also allow for the opposite effect, i.e. for $\lambda < q$.

To model demoralisation and encouragement, we assume that the higher the payoff $\pi(\bar{\bar{e}}^{t,p})$ that the agent believes the experimenter intends to give him (as compared to the neutral payoff $\hat{\pi}(\bar{\bar{e}}^{t,p})$), the more encouraged and the less resentful he is. Denoting by $v\{\pi_x[e(t,p),\bar{e}^{t,p}],\pi(\bar{\bar{e}}^{t,p}),\hat{\pi}(\bar{\bar{e}}^{t,p})\}$, the psychological payoff in the agent's utility derived from demoralisation and encouragement, a simple way to capture these motives is by assuming that:

$$\frac{\partial v\{\pi_x[e(t,p),\bar{e}^{t,p}],\pi(\bar{\bar{e}}^{t,p}),\hat{\pi}(\bar{\bar{e}}^{t,p})\}}{\partial \pi_x} = \pi(\bar{\bar{e}}^{t,p}) - \hat{\pi}(\bar{\bar{e}}^{t,p}). \tag{7}$$

For simplicity, we denote $\partial v\{\pi_x[e(t,p),\bar{e}^{t,p}],\pi(\bar{\bar{e}}^{t,p}),\hat{\pi}(\bar{\bar{e}}^{t,p})\}/\partial \pi_x$ by $v_{\pi_x}^{t,p}$. Since $\pi(\bar{\bar{e}}^{t,p})$ and $\hat{\pi}(\bar{\bar{e}}^{t,p}) \in [-\frac{1}{2},\frac{1}{4}]$, $v_{\pi_x}^{t,p} \in [-\frac{3}{4},\frac{3}{4}]$.[10]

Summarising, the belief-dependent utility of a reciprocal (t, p)-agent is the sum of the material and the psychological payoffs:

$$u^{t,p}[e(t,p),\bar{e}^{t,p},\bar{\bar{e}}^{t,p}] = t \times e(t,p) - e(t,p)^2 + v\{\pi_x[e(t,p),\bar{e}^{t,p}],\pi(\bar{\bar{e}}^{t,p}),\hat{\pi}(\bar{\bar{e}}^{t,p})\}. \tag{8}$$

This closes the description of our stylised randomised control trial with reciprocal agents. Next we analyse the impact of the procedure on the agents' behaviour.

2. Assignment Procedure Biases

In our context, an equilibrium in pure strategies is given by a profile of effort levels, such that the effort chosen by each type of agent maximises his utility for first-order and second-order beliefs that coincide with the equilibrium effort profile.[11] Denote with $e^*(t, p)$ the equilibrium effort level of a (t, p)-agent. Our first result concerns the existence of such an equilibrium in pure strategies.

Proposition 1. *The game exhibits a symmetric equilibrium in pure strategies. In every symmetric pure-strategy equilibrium effort levels are in the interior, i.e.* $0 < e^*(t, p) < 1$ *for all* t, p.

Proof. See Appendix.

A symmetric equilibrium in pure strategies always exists because the agents are homogenous and because for any given beliefs each agent's payoff function (including the psychological payoff) is convex.

Next we show that the effort levels of agents in both groups depend on whether the agents are assigned into the two groups through the private or the public randomisation procedure.

Proposition 2. *In every symmetric pure-strategy equilibrium, for every fraction* $q \in (0,1)$ *of agents assigned to the treatment group it holds:*

[10] Note that for $\lambda = \frac{1}{2}$ this specification of the psychological payoff is equivalent to the psychological payoff of the reciprocity models of Rabin (1993) and Dufwenberg and Kirchsteiger (2004).

[11] This equilibrium notion coincides with the equilibrium concept of Dufwenberg and Kirchsteiger (2004).

© 2015 Royal Economic Society.

$$e^*(1,d) > e^*(1,r) > e^*\left(\frac{1}{2},r\right) > e^*\left(\frac{1}{2},d\right).$$

Proof. See Appendix.

Proposition 2 shows that in policy experiments the treatment-induced differences in effort between the two groups are larger when the assignment into the two groups is done directly (i.e. through private randomisation) than when it is done using a public randomisation procedure. The effort is highest among privately chosen members of the treatment group and lowest among members of the privately assigned control group. The effort levels of agents allocated through a random assignment procedure are less extreme, with the effort of treatment-group agents still being higher than that of control-group agents. This result holds independent of the fraction of people that is assigned into the treatment group $q \in (0, 1)$.

The previous Proposition shows that randomisation procedures have an impact on the behaviour of agents in policy experiments. The key question then is: which procedure provides a correct prediction of the effect of a general introduction (scale-up) of the treatment, and under which circumstances does this occur?

In our setting, the effect of the programme scale-up to the entire population is the difference between the effort level of agents in the situation when the treatment is applied to everyone and the effort in the situation when the treatment is applied to nobody, i.e. between $q = 1$ and $q = 0$. We need to compare this difference to the difference in effort levels between agents in the treatment and control groups, under the two randomisation procedures.

PROPOSITION 3. *If the treatment is applied to everybody, i.e. if* $q = 1$, *then* $e^*(1, d) = e^*(1, r) = 1/2$. *In contrast, if the treatment is applied to nobody, i.e. if* $q = 0$, *then* $e^*(1/2, d) = e^*(1/2, r) = 1/4$.

Proof. See Appendix.

Proposition 3 shows that if nobody or everybody is chosen, the assignment procedure does not affect the effort and the effort chosen by an agent is as if he was motivated only by his material payoff. If all (none) of the agents are subject to the treatment, nobody feels encouraged (demotivated). Proposition 3 of course also reveals the true effect of the treatment (i.e. the difference between no and full introduction of the treatment).

The assignment through a private randomisation procedure *always* leads to an overestimation of the impact of the treatment, as the following Proposition shows.

PROPOSITION 4. *In every symmetric pure-strategy equilibrium, for every fraction* $q \in (0,1)$ *of agents assigned to the treatment group it holds:*

$$e^*(1,d) > \frac{1}{2} \quad \text{and} \quad e^*\left(\frac{1}{2},d\right) < \frac{1}{4}.$$

© 2015 Royal Economic Society.

Proof. See Appendix.

Under a private randomisation assignment, the untreated agents (i.e. the control group) are demotivated and hence their effort level is always smaller than the effort level realised when the entire population does not receive the treatment. On the other hand, the treated agents are encouraged when privately selected and hence their effort level is always larger than the one realised when the entire population receives the treatment. Therefore, any estimate of the effect of a general introduction of the treatment based on a policy experiment with private randomisation is biased upwards. A policy maker scaling up the programme on the basis of such an RCT faces the risk of introducing a non-effective programme to the entire population.

One might hope that with an explicit and credible randomisation procedure the treatment-induced differential effort in the policy experiment is the same as the one induced by a general introduction of the treatment. However, as the following Proposition shows, this need not be the case.

PROPOSITION 5.

(*i*) *For any* $\lambda \in (0, 1)$ *there exists at most one* q *such that* $e^*(1, r) - e^*(1/2, r) = 1/4$.
(*ii*) *If* $\lambda = q \in (0, 1)$, *then it holds in every symmetric pure-strategy equilibrium that* $e^*(1, r) - e^*(1/2, r) \neq 1/4$.

Proof. See Appendix.

Explicit randomisation does not solve the problem of the assignment procedure bias. Generically, the experimental results still do not provide a correct prediction of the impact of a general introduction of the treatment. There is no reason why the neutral payoff should equal the expected material payoff of an agent subject to explicit randomisation. Hence, even under a public randomisation the experimental results do not reflect the true benefits of a general introduction of the treatment.[12] This is true even for the natural case of $\lambda = q$ when agents have a 'rational' perception of how much they deserve the treatment.

While explicit randomisation does not completely solve the problem of a biased estimation of the true impact of the treatment, the following result shows that it certainly minimises its magnitude. Denote by b^p the bias generated by procedure p. It is the difference in effort levels between treatment and control group subjects for a given assignment procedure p minus the true effect of the treatment:

$$b^p = e(1, p) - e\left(\frac{1}{2}, p\right) - \frac{1}{4}. \tag{9}$$

Using this variable, we can state the following result:

PROPOSITION 6. *If* $\lambda = q \in (0, 1)$, *then it holds in every symmetric pure-strategy equilibrium that* $|b^d| > |b^r|$.

[12] In fact, with explicit randomisation even the sign of the prediction bias is unclear.

© 2015 Royal Economic Society.

Proof. See Appendix.

When agents have a 'rational' perception of how much they deserve the treatment, i.e. $\lambda = q$, the bias of the estimate of the true effect is always lower when subjects are assigned to the treatment and control group with the help of an unbiased and credible randomisation procedure relative to a direct appointment mechanism. This holds because the encouragement (demoralisation) of a treated (untreated) agent is lower if randomly selected than if directly appointed.

3. Discussion

The previous Sections show how Hawthorne and John Henry effects can be formalised and analysed by taking into account people's belief-dependent reciprocal preferences. In our model people's disappointment or elation is based on their belief about the experimenter's (un)kindness towards them. If they feel treated unkindly by the experimenter, they feel disappointment and resentment and, as a consequence, they are demotivated. On the other hand, if they feel kindly treated, they feel gratitude towards the experimenter and try to reciprocate by making the treatment a success.[13] Furthermore, the previous Sections highlight the impact of the choice of assignment procedure on the motivation and consequently on the effort levels of the subjects allocated into the control and treatment group. But how important are these effects in real-life RCTs in economics and other social sciences? Does resentful demoralisation and encouragement influence the behaviour of people in treatment and control groups as predicted by our theory?

There are several examples where this demoralisation effect played a key role for the results of policy experiments. One such study is the Birmingham Homeless Project (Schumacher *et al.*, 1994), aimed at homeless drug-users in Birmingham, Alabama. The randomly assigned subjects of the treatment group in this study received more frequent and therapeutically superior treatment, as compared to those in the control group. Schumacher *et al.* (1994, p. 42) note that an '11% increase in cocaine relapse rates for usual care clients [i.e. the control group, as compared to the general pre-treatment baseline level] was revealed'. They conclude, 'demoralisation represented a potential threat to the validity of this study [...] If the worsening of the usual care clients [control group] from baseline to the two-month follow-up point was related to demoralisation, there exists a potential for an overestimation of treatment effects of the enhanced care programme' (Schumacher *et al.*, 1994, p. 43–4).

Another example is the Baltimore Options Programme (Friedlander *et al.*, 1985), which was designed to increase the human capital and, hence, the employment possibilities of unemployed young welfare recipients in Baltimore Country. Half of the potential recipients were randomly assigned to the treatment group and half to the control group. The treatment group individuals in this RCT received tutoring and job

[13] There might of course be other sources of demotivation and elation that are unrelated to people's beliefs about the experimenter's (un)kindness. We abstract from these possible explanations in our analysis, since we want to present a unified theory based on one behavioural trait (i.e. reciprocity) that can jointly explain the Hawthorne and John Henry effects in RCTs.

© 2015 Royal Economic Society.

search training for one year. The control group members, aware of not having received the (desirable) treatment, performed worse in the outcome measure than they would have performed if the treatment group did not exist – leading to an overestimation of the effectiveness of the programme. Researchers found that the earnings of the treatment group increased by 16% but that the overall welfare claims of programme participants did not decrease. In line with Propositions 3 and 4 in the previous Section this implies that some of the control-group individuals in this study that would have normally moved out of welfare stayed longer on welfare because of the experiment.

Recent RCTs in development economics are characterised by factors making the presence of resentful demoralisation and encouragement particularly likely. First, in developing countries the economic value of inputs provided to subjects in the treatment groups is usually quite large (as compared, for instance, to their monthly incomes). Furthermore, in most RCTs the randomisation into the treatment and the control group is conducted privately (Bruhn and McKenzie, 2009). At the same time, subjects in the control group are typically aware of the existence of the treatment group and *vice versa*.[14] Consider, for example, the prominent randomised experiment of providing conditional cash transfers to poor families in Mexico, the PROGRESA/Opportunidades programme (see Levy (2006), and Parker *et al.* (2008) for a detailed survey of the various studies based on data from PROGRESA). The randomisation in the PROGRESA programme was privately done by Mexican government officials in charge of the RCT at the local level.[15] One of the concerns raised in the description of the randomisation in the first round of the programme (Behrman and Todd, 1999, p. 3) was that 'contamination could occur if families or individuals from control localities or other localities immigrate to treatment group localities in order to receive programme services. This would undermine the initial randomness of the samples, so it will be important to keep track of individuals leaving or entering the localities'. This indicates that the households in the control localities were well aware of the existence of the treatment localities and that the economic value of the programme benefits for them was substantial.[16] We are not aware of any study that explicitly discusses and analyses the presence or absence of Hawthorne and John Henry effects in the PROGRESA programme. However, it is interesting to note that in one of the recent studies (Angelucci and De Giorgi (2009, Table 1, p. 494) the growth in the observed difference in the outcome variable (food consumption) between the treatment and the

[14] Subjects might be completely unaware that they are part of an experimental study (i.e. it is a 'natural field experiment' in Levitt and List, 2009 taxonomy). However, they learn quickly that some households (or other units) receive certain benefits or special treatments, while others do not.

[15] Parker *et al.* (2008, p. 3980) note: 'Randomization was arguably an equitable method of assigning benefits in the context of limited resources (although this argument was *not* made publicly in Mexico at the time) . . . There is unfortunately little written evidence on how precisely the randomisation was done . . . The lack of documentation by government officials may reflect their perception of the controversial nature of carrying out an evaluation with an experimental design. In fact when the results of the initial evaluation studies were made public in 2000, a number of Mexico City newspapers ran articles criticising the "unethical nature" of the evaluation'.

[16] Over time the size of the programme increased substantially. As Parker *et al.* (2008, p. 3981) note: 'This rapid growth created a scenario where many of the original control communities began literally to become "surrounded" by communities (presumably similar to themselves) receiving programme benefits'.

© 2015 Royal Economic Society.

control group is at least in part driven by a *reduction* in the outcome in the control-group households which is a clear indication of the John Henry effect.

As mentioned by the authors themselves, a Hawthorne effect was likely to be present in another well-known study in development economics, namely Banerjee *et al.* (2007). This study analyses the effect of remedial education on learning outcomes of poor children in Indian urban schools. In this study, the subject pool consisted of 3rd and 4th graders in 98 schools. In (privately randomly selected) half of the schools, fourth-grade children received remedial education to teach those of them lagging behind in basic literacy and numeracy skills. In the other half, remedial education was provided to third-graders. The existence of some treatment-group subjects in every school, of course, implied that subjects could easily become aware of the existence of the treatment and control groups. The authors find that in the short-run, providing remedial education increased average test scores of children in treatment groups (as compared to those in control groups) by 0.28 standard deviation. However, one year after the intervention, initial gains faded to about 0.10 standard deviation. The authors say that one possible explanation for a large short-run and the much smaller long-run impacts is the Hawthorne effect: 'Children exposed to the *balsakhi* or to computers [the two components of the remedial education intervention] may feel grateful and compelled to exert their best effort while taking the test' (Banerjee *et al.*, 2007, p. 1257). This is precisely what our Proposition 4 in the previous Section predicts.

The presence of the same problem can be observed in an RCT aiming at studying the effect of new agricultural technologies in rural Tanzania (Bulte *et al.*, 2014). In one sub-experiment of this study the experimenters allocated the more advanced technology (i.e. better-quality seeds) to the treatment group using the standard private-randomisation RCT design so that the subjects knew to which group they belonged. In the second sub-experiment the subjects did not know whether they belonged to the treatment or the control group that received lower-quality traditional seeds. The authors find that harvests were virtually equal for the subjects who knew that they received modern seeds and for those who did not know what type of seeds they received (regardless of the type of seeds that they actually received). However subjects who *knew* that they received the traditional seeds did much worse. In fact, in the first sub-experiment, the subjects in the control group were likely to become demotivated and provided a relatively low level of effort in taking care of the harvest. In the second sub-experiment, all the subjects provided a relatively high level of effort. Thus, the difference between the outcomes of the control group subjects in the first and the second sub-experiment indicates differences in effort levels likely due to demotivation.[17,18]

[17] An alternative explanation for this result is an the effort response that is nonlinear in (subjective) beliefs about treatment. Suppose that the effort provision by people who believed they are treated with probability 0.5 is very close to that of people sure of being in the treatment group. Furthermore assume that this effort is much larger than that of people who are sure that they are in the control group. Then, the observed results can emerge even without resentful demoralisation or encouragement. However, the authors explain that during the experiment the subjects (in all groups) were clearly informed that the improved type of seeds were more productive than the traditional type. Therefore it is likely that those who knew that they received the traditional type were unhappy with the allocation and were demotivated.

[18] Beyond development economics, several experimental studies have considered the potential impact of the Hawthorne effect on the validity of their results. See Krueger (1999) for a classic study in public economics and Harris (1985) for a critical analysis of two experiments in health economics.

© 2015 Royal Economic Society.

What are the implications of our analysis for the design of future policy experiments? First, given that our analysis indicates that the assignment procedure into the treatment and control groups matters, any experimental study should clearly indicate how exactly the randomisation was carried out and discuss, wherever possible, how subjects perceived belonging to one of the two groups.[19] Second, prior to the experiment, the relevant outcome of all the subjects should be measured before the RCT, and changes in the outcomes for the treatment and the control group subjects should be reported (and not just the difference in post-intervention levels). Furthermore, as suggested by Duflo *et al.*, (2008), data on the behaviour of the subjects in both groups should be collected in the long run. If the Hawthorne and John Henry effects are relatively short-lived, the true effect of the treatment might be identified by the long-run data. Finally, we have shown that the assignment procedure bias is minimised by public randomisation. If possible, public lotteries should be used to allocated subjects into the two groups. Another possibility often used in the RCTs in developing countries (as, for instance, in the PROGRESA/Oportunidades programme) is a randomised phasing-in of the treatment (in which control-group subjects also receive the treatment but at a later date). This method, however, has its own downsides: for example, control-group subjects might start changing their behaviour early, in expectation of future treatment (as has been suggested by Duflo *et al.* (2008) for microfinance experiments). In this case, the obtained estimates of the treatment effect would also be biased.

4. Conclusion

In this article, we analyse encouragement and resentful demoralisation (two expressions of the Hawthorne and John Henry effects), their common behavioural root, and their impact on the validity of policy experiments. We show that if agents are prone to demoralisation and encouragement, the way in which experimenters assign them into the treatment and control groups influences their behaviour. Thus, the size of the estimated treatment effect depends on the assignment procedure. If agents are assigned directly into the treatment and control group (i.e. via a private randomisation), or if agents believe that they are assigned directly, the experimentally observed treatment effect is always larger than the effect of a general introduction of the treatment. This assignment procedure bias is always smaller for a credible (explicit/public) randomisation procedure.

Our analysis concentrates on the effects of reciprocity and, hence, demoralisation and encouragement. There are other belief-dependent motives like guilt (Charness and Dufwenberg, 2006; Battigalli and Dufwenberg, 2007) or disappointment (Ruffle, 1999) that have been found to affect agents' behaviour. Exploring the impact of these effects on the validity of RCTs is left to future research.

[19] It is discomforting that for the most well-known policy experiment in development economics, the PROGRESA programme, no written document describes how the randomisation was carried out.

© 2015 Royal Economic Society.

Appendix A

A.1. *Proof of Proposition* 1

Recall that $\pi(\bar{\bar{e}}^{t,p})$ and $\hat{\pi}(\bar{\bar{e}}^{t,p})$ depend only on the agent's second-order beliefs about the effort (and not on the effort level itself) and that $\partial\pi_x[e(t,p),\bar{e}^{t,p}]/\partial e(t,p) = t$. Hence:

$$\frac{\partial u^{t,p}[e(t,p),\bar{e}^{t,p},\bar{\bar{e}}^{t,p}]}{\partial e(t,p)} = t(1+v_{\pi x}^{t,p}) - 2e(t,p), \tag{A.1}$$

$$\frac{\partial^2 u^{t,p}[e(t,p),\bar{e}^{t,p},\bar{\bar{e}}^{t,p}]}{\partial e(t,p)^2} = \frac{\partial^2 v\{\pi_x[e(t,p),\bar{e}^{t,p}],\pi(\bar{\bar{e}}^{t,p}),\hat{\pi}(\bar{\bar{e}}^{t,p})\}}{(\partial\pi_x)^2}\,t^2 - 2. \tag{A.2}$$

Since:

$$\begin{aligned} \frac{\partial^2 v\{\pi_x[e(t,p),\bar{e}^{t,p}],\pi(\bar{\bar{e}}^{t,p}),\hat{\pi}(\bar{\bar{e}}^{t,p})\}}{(\partial\pi_x)^2} &= 0, \\ \frac{\partial^2 u^{t,p}[e(t,p),\bar{e}^{t,p},\bar{\bar{e}}^{t,p}]}{\partial e(t,p)^2} &< 0 \text{ for all } t,p. \end{aligned} \tag{A.3}$$

Because $|v_{\pi x}^{t,p}| \leq 3/4$, it is easy to check that:

$$\begin{aligned} \left.\frac{\partial u^{t,p}[e(t,p),\bar{e}^{t,p},\bar{\bar{e}}^{t,p}]}{\partial e(t,p)}\right|_{e(t,p)=0} &> 0 \text{ for all } t,p, \\ \left.\frac{\partial u^{t,p}[e(t,p),\bar{e}^{t,p},\bar{\bar{e}}^{t,p}]}{\partial e(t,p)}\right|_{e(t,p)=1} &< 0 \text{ for all } t,p. \end{aligned} \tag{A.4}$$

Because of (A.3) and (A.4), each of the equations

$$\frac{\partial u^{t,p}[e(t,p),\bar{e}^{t,p},\bar{\bar{e}}^{t,p}]}{\partial e(t,p)} = 0 \tag{A.5}$$

has a unique interior solution for each t, p for any first and second-order belief $\bar{e}^{t,p}, \bar{\bar{e}}^{t,p}$. These solutions characterise the optimal effort choices of all types of agents for given first-order and second-order beliefs. In equilibrium, the beliefs of first-order and second-order have to be the same, i.e. $\bar{e}^{t,p} = \bar{\bar{e}}^{t,p}$ for all t, p. The solution of (A.5) can be rewritten as a function:

$$e_{opt}^{t,p} : [0,1]^4 \to [0,1]^4,$$

with $e_{opt}^{t,p}(\bar{e}^{t,p})$ being the optimal effort choice of an (t, p)-agent who holds the same first-order and second-order beliefs $\bar{e}^{t,p} = \bar{\bar{e}}^{t,p}$. Since $u^{t,p}[e(t,p),\bar{e}^{t,p},\bar{\bar{e}}^{t,p}]$ is twice continuously differentiable, $e_{opt}^{t,p}$ is also continuous. Brower's fixed-point theorem guarantees the existence of a fixed point:

$$\exists e^* \in [0,1]^4 : e_{opt}^{t,p}(e^*) = e^*(t,p) \text{ for all } t,p.$$

The effort levels characterised by this fixed point maximise the agents' utilities for first-order and second-order beliefs which coincide with the utility maximising effort levels, i.e. for correct beliefs. Hence, e^* fulfils the conditions for an equilibrium.

© 2015 Royal Economic Society.

A.2. *Proof of Proposition* 2

By Proposition 1, the equilibrium effort levels are in the interior. Hence, they are fully characterised by the first-order conditions (FOCs):

$$1 - 2e(1, d) + v_{\pi x}^{1,d} = 0, \tag{A.6}$$

$$\frac{1}{2} - 2e\left(\frac{1}{2}, d\right) + v_{\pi x}^{\frac{1}{2},d}\frac{1}{2} = 0, \tag{A.7}$$

$$1 - 2e(1, r) + v_{\pi x}^{1,r} = 0, \tag{A.8}$$

$$\frac{1}{2} - 2e\left(\frac{1}{2}, r\right) + v_{\pi x}^{\frac{1}{2},r}\frac{1}{2} = 0. \tag{A.9}$$

In equilibrium, the beliefs have to be correct. The FOCs hold with $\bar{\bar{e}}^{t,p}(t', p') = \bar{e}^{t,p}(t', p') = e(t', p')$.

To prove the Proposition, we first show that $e^*(1, r) > e^*(1/2, r)$. Since in equilibrium $\bar{\bar{e}}^{\frac{1}{2},r}(t', p') = \bar{\bar{e}}^{1,r}(t', p') = e(t', p')$, $\bar{\bar{\pi}}_a^{1,r}(\bar{\bar{e}}^{1,r}) = \bar{\bar{\pi}}_a^{\frac{1}{2},r}(\bar{\bar{e}}^{\frac{1}{2},r})$. Because of this equality, $v_{\pi x}^{1,r} = v_{\pi x}^{\frac{1}{2},r}$. Using this and comparing the FOCs (A.8) and (A.9) reveal that $e^*(1, r) = 2e^*(1/2, r) > e^*(1/2, r)$.

Second, we prove that:

$$e^*(1, r) - e^*(1, r)^2 > \frac{1}{2}e^*\left(\frac{1}{2}, r\right) - e^*(\frac{1}{2}, r)^2. \tag{A.10}$$

Inserting $e^*(1, r) = 2e^*(1/2, r)$ and rearranging terms, (A.10) becomes:

$$\frac{3}{4}[e^*(1, r) - e^*(1, r)^2] > 0,$$

which holds for any $e^*(1, r) \in (0, 1)$.

Third, it has to be shown that $e^*(1, d) > e^*(1, r)$. Because of (5), (7) and (A.10) it is true that:

$$\begin{aligned} v_{\pi x}^{1,d} - v_{\pi x}^{1,r} &= e(1, d) - e(1, d)^2 - q[e(1, r) - e(1, r)^2] - (1 - q)\left[\frac{1}{2}e\left(\frac{1}{2}, r\right) - e\left(\frac{1}{2}, r\right)^2\right] \\ &> e(1, d) - e(1, d)^2 - e(1, r) + e(1, r)^2. \end{aligned}$$

Comparing (A.6)–(A.8), one sees that:

$$v_{\pi x}^{1,d} - v_{\pi x}^{1,r} = 2[e(1, d) - e(1, r)], \tag{A.11}$$

implying that

$$e(1, d) - e(1, r) > -e(1, d)^2 + e(1, r)^2. \tag{A.12}$$

However, this condition can only hold for $e^*(1, d) > e^*(1, r)$.

Finally, it remains to show that $e^*(1/2, r) > e^*(1/2, d)$. Because of (5), (7) and (19), it holds that:

$$\begin{aligned} v_{\pi x}^{\frac{1}{2},r} - v_{\pi x}^{\frac{1}{2},d} &= q[e(1, r) - e(1, r)^2] + (1 - q)\left[\frac{1}{2}e\left(\frac{1}{2}, r\right) - e\left(\frac{1}{2}, r\right)^2\right] - \frac{1}{2}e\left(\frac{1}{2}, d\right) + e\left(\frac{1}{2}, d\right)^2 \\ &> \frac{1}{2}\left[e\left(\frac{1}{2}, r\right) - e\left(\frac{1}{2}, d\right)\right] - e\left(\frac{1}{2}, r\right)^2 + e\left(\frac{1}{2}, d\right)^2. \end{aligned}$$

© 2015 Royal Economic Society.

Comparing (A.7) to (A.9), one gets:

$$v_{\pi x}^{\frac{1}{2},r} - v_{\pi x}^{\frac{1}{2},d} = 4\left[e\left(\frac{1}{2},r\right) - e\left(\frac{1}{2},d\right)\right],$$

implying that

$$\frac{7}{2}\left[e\left(\frac{1}{2},r\right) - e\left(\frac{1}{2},d\right)\right] > -e\left(\frac{1}{2},r\right)^2 + e\left(\frac{1}{2},d\right)^2.$$

However, this condition can only hold for $e^*(1/2, r) > e^*(1/2, d)$.

A.3. *Proof of Proposition* 3

(*i*) $q = 1$ implies that $\lambda = 1$. Therefore, $\pi(\bar{e}^{1,d}) = \hat{\pi}(\bar{e}^{1,d})$ and $v_{\pi x}^{1,d} = 0$. From (A.6) it follows that $e^*(1, d) = 1/2$. Since the beliefs have to be correct in equilibrium, we get that $\hat{\pi}(\bar{e}^{1,r}) = 1/4$. By substituting into (A.8) we get:

$$1 - 2e(1,r) + \left[e(1,r) - e(1,r)^2 - \frac{1}{4}\right] = 0, \tag{A.13}$$

given that the beliefs have to be correct. The unique solution to (A.13) is $e^*(1, r) = 1/2$.

(*ii*) $q = 0$ implies that $\lambda = 0$. Therefore, $\pi(\bar{e}^{\frac{1}{2},d}) = \hat{\pi}(\bar{e}^{\frac{1}{2},d})$ and $v_{\pi x}^{1,d} = 0$. From (A.7) it follows that $e^*(1, d) = 1/4$. Since the beliefs have to be correct in equilibrium, we get $\hat{\pi}(\bar{e}^{1,r}) = 1/16$. By substituting into (A.9) we get:

$$\frac{1}{2} - 2e\left(\frac{1}{2},r\right) + \frac{1}{2}\left[\frac{1}{2}e\left(\frac{1}{2},r\right) - e\left(\frac{1}{2},r\right)^2 - \frac{1}{16}\right] = 0, \tag{A.14}$$

given that the beliefs have to be correct. The unique solution to (A.14) is $e^*(1/2, r) = 1/4$.

A.4. *Proof of Proposition* 4

We first show that in equilibrium $v_{\pi x}^{1,d} > 0 > v_{\pi x}^{\frac{1}{2},d}$. Inserting (5) and (6) into (7) gives:

$$v_{\pi x}^{1,d} = (1-\lambda)\left[e(1,d) - e(1,d)^2 - \frac{1}{2}e\left(\frac{1}{2},d\right) + e\left(\frac{1}{2},d\right)^2\right], \tag{A.15}$$

$$v_{\pi x}^{\frac{1}{2},d} = -\lambda\left[e(1,d) - e(1,d)^2 - \frac{1}{2}e\left(\frac{1}{2},d\right) + e\left(\frac{1}{2},d\right)^2\right].$$

Both equations together can only hold for either $v_{\pi x}^{1,d} = v_{\pi x}^{\frac{1}{2},d} = 0$ or for $v_{\pi x}^{1,d}$ and $v_{\pi x}^{\frac{1}{2},d}$ having opposite signs.

Take first the case of $v_{\pi x}^{1,d} = v_{\pi x}^{\frac{1}{2},d} = 0$. In this case, the equilibrium effort levels would be $1/2$ and $1/4$, respectively (see FOCs (A.6) and (A.7)). Inserting these values and (5) and (6) into (7), one obtains that $v_{\pi x}^{1,d} > 0 > v_{\pi x}^{\frac{1}{2},d}$ – a contradiction.

Hence, $v_{\pi x}^{1,d}$ and $v_{\pi x}^{\frac{1}{2},d}$ must have opposite signs. Assume that $v_{\pi x}^{1,d} < 0 < v_{\pi x}^{\frac{1}{2},d}$. This inequality together with the FOCs (A.6) and (A.7) implies that $e(1, d) < 1/2$ and $e(1/2, d) > 1/4$. Since $e(1, d) > e(1/2, d)$, this implies that $e(t, d) \in (1/4, 1/2)$ for $t = 1,1/2$.

© 2015 Royal Economic Society.

Because of (A.15) and $v_{\pi x}^{1,d} < 0 < v_{\pi x}^{\frac{1}{2},d}$,

$$-e(1,d) + e(1,d)^2 + \frac{1}{2}e\left(\frac{1}{2},d\right) - e\left(\frac{1}{2},d\right)^2 = -v_{\pi x}^{1,d} + v_{\pi x}^{\frac{1}{2},d} > 0. \tag{A.16}$$

For $e(t, d) \in (1/4, 1/2)$ the left-hand side of (A.16) is decreasing in $e(1, d)$ and $e(1/2, d)$. However, even for the limit case of $e(1, d) = e(1/2, d) = 1/4$ the left hand side of (A.16) is $-1/8$. Hence (A.16) cannot hold and $v_{\pi x}^{1,d} < 0 < v_{\pi x}^{\frac{1}{2},d}$ is not possible in equilibrium. Therefore, $v_{\pi x}^{1,d} > 0 > v_{\pi x}^{\frac{1}{2},d}$. This and (A.15) also imply that $e^*(1, d) - e^*(1, d)^2 > 1/2e^*(1/2, d) - e^*(1/2, d)^2$–the material payoff from getting a treatment is larger than from not getting a treatment, if the selection is done directly.

Recall that $v_{\pi x}^{t,p} \in [-3/4, 3/4]$. Hence, $v_{\pi x}^{1,d} \in (0, 3/4]$ and $v_{\pi x}^{\frac{1}{2},d} \in [-3/4, 0)$. Using this and the FOCs (A.6) and (A.7) one immediately gets that $e^*(1, d) \in (1/2, 7/8]$ and that $e^*(1/2, d) \in [1/16, 1/4)$.

A.5. *Proof of Proposition* 5

(*i*) Subtracting (A.9) from (A.8) reveals that $v_{\pi x}^{1,r} - 1/2 v_{\pi x}^{\frac{1}{2},r} = 0$, whenever in equilibrium $e(1, r) - e(1/2, r) = 1/4$. Since $v_{\pi x}^{1,r} = v_{\pi x}^{1/2,r}$, this can only hold for $v_{\pi x}^{1,r} = v_{\pi x}^{1/2,r} = 0$. Hence, $e(1, r) = 1/2$, $e(1/2, r) = 1/4$ in equilibrium if the difference in equilibrium effort is $1/4$.

In equilibrium, the beliefs have to be correct. From this, $v_{\pi x}^{1,r} = v_{\pi x}^{1/2,r} = 0$, and $e(1, r) = 1/2$, $e(1/2, r) = 1/4$, we get that in equilibrium the neutral payoff must be given by:

$$\hat{\pi} = \frac{3q+1}{16}. \tag{A.17}$$

Using the definition of $\hat{\pi}$, (A.17), and again the fact that the equilibrium beliefs are correct, we get:

$$\frac{3q+1}{16} = \lambda\pi(1,d) + (1-\lambda)\pi\left(\frac{1}{2},d\right). \tag{A.18}$$

If in equilibrium $e(1, r) - e(1/2, r) = 1/4$, then (A.18) has to hold. Recall that $\pi(1, d)$ and $\pi(1/2, d)$ are determined by the joint solution of the FOCs (A.6) and (A.7). Since $v_{\pi x}^{t,d}$ is independent of q, $\pi(1, d)$ and $\pi(1/2,d)$ do not depend on q. Hence the right-hand side of (A.18) is independent of q, whereas the left-hand side is strictly increasing in q. Hence, for any given $\lambda \in (0, 1)$ there exists at most one q such that $e_1^* - e_0^* = 1/4$.

(*ii*) Inserting (A.17) into (7) and (A.6) leads to:

$$1 - 2e(1,d) + \left[e(1,d) - e(1,d)^2 - \frac{3q+1}{16}\right] = 0.$$

By solving this equation one gets:

$$e(1,d) = \frac{-2 + \sqrt{19 - 3q}}{4}. \tag{A.19}$$

Inserting (A.17) into (7) and (A.7) leads to:

$$\frac{1}{2} - 2e\left(\frac{1}{2},d\right) + \left[\frac{1}{2}e\left(\frac{1}{2},d\right) - e\left(\frac{1}{2},d\right)^2 - \frac{3q+1}{16}\right]\frac{1}{2} = 0.$$

© 2015 Royal Economic Society.

By solving this equation one gets:

$$e\left(\frac{1}{2}, d\right) = \frac{-7 + \sqrt{64 - 3q}}{4}. \tag{A.20}$$

Given that $\lambda = q$ and because of (A.20) and (A.19), (A.18) becomes:

$$\frac{3q+1}{16} = q\left[\frac{-2 + \sqrt{19 - 3q)}}{4} - \left(\frac{-2 + \sqrt{19 - 3q)}}{4}\right)^2\right] + (1-q)\left[\frac{1}{2}\frac{-7 + \sqrt{64 - 3q}}{4} - \left(\frac{-7 + \sqrt{64 - 3q}}{4}\right)^2\right], \tag{A.21}$$

leading to

$$0 = 96q + 8q\sqrt{19 - 3q} - 16q\sqrt{64 - 3q} + 16\sqrt{64 - 3q} - 128. \tag{A.22}$$

For any $q \in (0, 1)$, the right-hand side of (A.22) is strictly larger than zero. This equation holds only for the limit cases $q = 1$ and $q = 0$.

A.6. *Proof of Proposition* 6

Due to Proposition 2, $b^d > 0$. Hence, we have to distinguish between two cases:

(*i*) $b^r \geq 0$. In this case Proposition 4 implies that $|b^d| > |b^d|$; and
(*ii*) $b^r < 0$.

From the definition (9), the first order conditions (A.6)–(A.9) and taking into account that $v_{\pi x}^{1,r} = v_{\pi x}^{\frac{1}{2},r} = v_{\pi x}^{r}$ we get:

$$b^d = \frac{2v_{\pi x}^{1,d} - v_{\pi x}^{\frac{1}{2},d}}{4}, \text{ and}$$
$$b^r = \frac{v_{\pi x}^{r}}{4}.$$

Because $b^r < 0$, the Proposition holds if:

$$b^d + b^r = \frac{2v_{\pi x}^{1,d} - v_{\pi x}^{\frac{1}{2},d} + v_{\pi x}^{r}}{4} > 0. \tag{A.23}$$

Because of (5), (7), and (6), and since in equilibrium expectations are correct, we get:

$$v_{\pi x}^{1,d} = (1-\lambda)\left\{\pi[1, e(1,d)] - \pi\left[\frac{1}{2}, e\left(\frac{1}{2}, d\right)\right]\right\},$$

$$v_{\pi x}^{\frac{1}{2},d} = \lambda\left\{-\pi[1, e(1,d)] + \pi\left[\frac{1}{2}, e\left(\frac{1}{2}, d\right)\right]\right\},$$

$$v_{\pi x}^{r} = \lambda\{\pi[1, e(1,r)] - \pi[1, e(1,d)]\} + (1-\lambda)\left\{\pi\left[\frac{1}{2}, e\left(\frac{1}{2}, r\right)\right] - \pi\left[\frac{1}{2}, e\left(\frac{1}{2}, d\right)\right]\right\}.$$

© 2015 Royal Economic Society.

Inserting this, condition (A.23), the condition for the Proposition to hold becomes:

$$2(1-\lambda)\left\{\pi[1,e(1,d)]-\pi\left[\frac{1}{2},e\left(\frac{1}{2},d\right)\right]\right\}+\lambda\pi[1,e(1,r)]$$
$$+(1-\lambda)\pi\left[\frac{1}{2},e\left(\frac{1}{2},r\right)\right]-\pi\left[\frac{1}{2},e\left(\frac{1}{2},d\right)\right]>0. \quad \text{(A.24)}$$

Recall that the agent gets the maximum material payoff when directly appointed to the treatment group. Furthermore, the expected material payoff of random assignment is higher than the material payoff of direct appointment to the control group. Hence condition (A.24) holds.

THEMA, University of Cergy-Pontoise and CRED, University of Namur
ECARES, Université Libre de Bruxelles, ECORE, CEPR, CESifo, and Vienna Center for Experimental Economics
University of Copenhagen

Submitted: 9 January 2014
Accepted: 26 May 2015

References

Abbring, J. and Heckman, J. (2007). 'Econometric evaluation of social programs, part III: distributional treatment effects, dynamic treatment effects, dynamic discrete choice, and general equilibrium policy evaluation', in (J. Heckman, and E. Leamer, eds.), *Handbook of Econometrics*, vol. 6, pp. 5145–5303, Amsterdam: Elsevier.

Angelucci, M. and De Giorgi, G. (2009). 'Indirect effects of an aid program: how do cash transfers affect ineligibles' consumption?', *American Economic Review*, vol. 99(1), pp. 486–508.

Angrist, J., Bettinger, E., Bloom, E., King, E. and Kremer, M. (2002). 'Vouchers for private schooling in Colombia: evidence from a randomized natural experiment', *American Economic Review*, vol. 92(5), pp. 1535–58.

Angrist, J., Bettinger, E. and Kremer, M. (2006). 'Long-term educational consequences of secondary school vouchers: evidence from administrative records in Colombia', *American Economic Review*, vol. 96(3), pp. 847–62.

Banerjee, A., Cole, S., Duflo, E. and Linden, L. (2007). 'Remedying education: evidence from two randomized experiments in India', *Quarterly Journal of Economics*, vol. 122(3), pp. 1235–64.

Banerjee, A., Duflo, E., Glennerster, R. and Kinnan, C. (2010). 'The miracle of microfinance: evidence from a randomized evaluation', Working Paper, Department of Economics, MIT.

Battigalli, P. and Dufwenberg, M. (2007). 'Guilt in games', *American Economic Review, Papers and Proceedings*, vol. 97(2), pp. 170–6.

Battigalli, P. and Dufwenberg, M. (2009). 'Dynamic psychological games', *Journal of Economic Theory*, vol. 144 (1), pp. 1–35.

Behrman, J. and Todd, P. (1999). *Randomness in the Experimental Sample of Progresa (Education, Health, and Nutrition Program)*, Washington, DC: International Food Policy Research Institute.

Bruhn, M. and McKenzie, D. (2009). 'In pursuit of balance: randomization in practice in development field experiments', *American Economic Journal: Applied Economics*, vol. 1(4), pp. 200–32.

Bulte, E., Beekman, G., Di Falco, S., Hella, J. and Lei, P. (2014). 'Behavioral responses and the impact of new agricultural technologies: evidence from a double-blind experiment in Tanzania', *American Journal of Agricultural Economics*, vol. 96(3), pp. 813–30.

Card, D. and Hyslop, D. (2005). 'Estimating the effects of a time-limited earnings subsidy for welfare-leavers', *Econometrica*, vol. 73(6), pp. 1723–70.

Card, D. and Robins, P. (2005). 'How important are "entry effects" in financial incentive programs for welfare recipients? Experimental evidence from the self-sufficiency project', *Journal of Econometrics*, vol. 125(1–2), pp. 113–39.

Charness, G. and Dufwenberg, M. (2006). 'Promises and partnership', *Econometrica*, vol. 74(6), pp. 1579–601.

Chassang, S., Padró-i-Miquel, G. and Snowberg, E. (2012). 'Selective trials: a principal-agent approach to randomized controlled experiments', *American Economic Review*, vol. 102(4), pp. 1279–309.

© 2015 Royal Economic Society.

Cook, T. and Campbell, D. (1979). *Quasi-experimentation: Design and Analysis Issues for Field Settings*, Boston, MA: Houghton Mifflin.
De Mel, S., McKenzie, D. and Woodruff, C. (2008). 'Returns to capital in microenterprises: evidence from a field experiment', *Quarterly Journal of Economics*, vol. 123(4), pp. 1329–72.
Duflo, E. (2004). '*Scaling Up and Evaluation*', Annual World Bank Conference on Development Economics, Washington, DC: The World Bank.
Duflo, E., Gale, W., Liebman, J., Orszag, P. and Saez, E. (2006). 'Saving incentives for low- and middle-income families: evidence from a field experiment with H & R block', *Quarterly Journal of Economics*, vol. 121(4), pp. 1311–46.
Duflo, E., Glennerster, R. and Kremer, M. (2008). 'Using randomization in development economics research: a toolkit', In (T.P. Schultz and J.A. Strauss, eds.) *Handbook of Development Economics*, vol. 4, pp. 3895–962. Amsterdam: Elsevier.
Dufwenberg, M. and Kirchsteiger, G. (2004). 'A theory of sequential reciprocity', *Games and Economic Behavior*, vol. 47(2), pp. 268–98.
Ferraz, C. and Finan, F. (2008). 'Exposing corrupt politicians: the effects of Brazil's publicly released audits on electoral outcomes', *Quarterly Journal of Economics*, vol. 123(2), pp. 703–45.
Fetterman, D. (1982). 'Ibsen's baths: reactivity and insensitivity (a misapplication of the treatment-control design in a national evaluation)', *Educational Evaluation and Policy Analysis*, vol. 4(3), pp. 261–79.
Friedlander, D., Hoetz, G., Long, D. and Quint, J. (1985). *Maryland: Final Report on the Employment Initiatives Evaluation*, New York, NY: MDRC.
Geanakoplos, J., Pearce, D. and Stacchetti, E. (1989). 'Psychological games and sequential rationality', *Games and Economic Behavior*, vol. 1(1), pp. 60–79.
Gertler, P. (2004). 'Do conditional cash transfers improve child health? Evidence from PROGRESA's controlled randomized experiment', *American Economic Review*, vol. 94(2), pp. 336–41.
Harris, J. (1985). 'Macroexperiments versus microexperiments for health policy', in(J. Hausmann, and D. Wise, eds.), *Social Experimentation*, pp. 145–185, Chicago, IL: University of Chicago Press.
Heckman, J. (1991) 'Randomization and social policy evaluation', Working Paper 107, National Bureau of Economic Research Technical.
Heckman, J. and Vytlacil, E. (2007*a*), 'Econometric evaluation of social programs, part I: causal models, structural models and econometric policy evaluation', in(J. Heckman, and E. Leamer eds.), *Handbook of Econometrics*, vol. 6, pp. 4779–4874, Amsterdam: Elsevier.
Heckman, J. and Vytlacil, E. (2007*b*), 'Econometric evaluation of social programs, part II: using the marginal treatment effects to organize alternative econometric estimators to evaluate social programs, and to forecast their effects in new environments', in (J. Heckman, and E. Leamer, eds.), *Handbook of Econometrics*, vol. 6, pp. 4875–5143, Amsterdam: Elsevier.
Hoorens, V. (1993). 'Self-enhancement and superiority biases in social comparison', *European Review of Social Psychology*, vol. 4(1), pp. 113–39.
Imbens, G. and Wooldridge, J. (2009). 'Recent developments in the econometrics of program evaluation', *Journal of Economic Literature*, vol. 47(1), pp. 5–86.
Krueger, A. (1999). 'Experimental estimates of education production functions', *Quarterly Journal of Economics*, vol. 114(2), pp. 497–532.
Levitt, S. and List, J. (2009). 'Field experiments in economics: the past, the present, and the future', *European Economic Review*, vol. 53(1), pp. 1–18.
Levitt, S. and List, J. (2011). 'Was there really a Hawthorne effect at the Hawthorne plant? An analysis of the original illumination experiments', *American Economic Journal: Applied Economics*, vol. 3(1), pp. 224–38.
Levy, S. (2006). *Progress against Poverty: Sustaining Mexico's Progresa-Oportunidades Program*, Washington, DC: Brookings Institution Press.
Malani, A. (2006). 'Identifying placebo effects with data from clinical trials', *Journal of Political Economy*, vol. 114(2), pp. 236–56.
Michalopoulos, C., Robins, P. and Card, D. (2005). 'When financial work incentives pay for themselves: evidence from a randomized social experiment for welfare recipients', *Journal of Public Economics*, vol. 89 (1), pp. 5–29.
Ongena, S. (2009). 'Resentful demoralization', in (B. Everitt and D. Howel, eds), *Encyclopedia of Statistics in Behavioral Science*, vol. 4, pp. 1744–1746, Chichester, UK: Wiley.
Parker, S., Rubalcava, L. and Teruel, G. (2008). 'Evaluating conditional schooling and health programs', in (T.P. Schultz and J. Strauss eds.), *Handbook of Development Economics*, vol. 4, pp. 3963–4035, North-Holland: Elsevier.
Philipson, T. and Hedges, L. (1998). 'Subject evaluation in social experiments', *Econometrica*, vol. 66(2), pp. 381–408.
Rabin, M. (1993). 'Incorporating fairness into game theory and economics', *American Economic Review*, vol. 83 (4), pp. 1281–302.
Ruffle, B. (1999). 'Gift giving with emotions', *Journal of Economic Behavior & Organization*, vol. 39(4), pp. 399–420.

© 2015 Royal Economic Society.

Schultz, T.P. (2004). 'School subsidies for the poor: evaluating the Mexican Progresa poverty program', *Journal of Development Economics*, vol. 74(1), pp. 199–250.
Schumacher, J., Milby, J., Raczynski, J., Engle, M., Caldwell, E. and Carr, J. (1994). 'Demoralization and threats to validity in Birmingham's Homeless Project', in (K. Conrad ed.), *Critically Evaluating the Role of Experiments*, pp. 41–44, San Francisco, CA: Jossey-Bass.
Sebald, A. (2010). 'Attribution and reciprocity', *Games and Economic Behavior*, vol. 68(1), pp. 339–52.
Sobel, J. (2005). 'Interdependent preferences and reciprocity', *Journal of Economic Literature*, vol. 43(2), pp. 392–436.

© 2015 Royal Economic Society.