# Assignment procedure biases in randomized policy experiments*

Gani Aldashev[†], Georg Kirchsteiger[‡] and Alexander Sebald[§]

September 2, 2013

**Abstract**

Randomized controlled trials (RCT) have gained ground as dominant tool for studying policy interventions in many fields of applied economics. We theoretically analyze feelings of encouragement and resentful demoralization in RCTs and show that these might be rooted in the same behavioral trait - people's propensity to act reciprocally. Interestingly, when people are motivated by reciprocity, the choice of assignment procedure influences the RCTs' findings. We show that even credible and explicit randomization procedures do not guarantee an unbiased prediction of the impact of policy interventions, however they minimize any bias relative to other less transparent assignment procedures.

**Keywords:** Randomized controlled trials, Policy experiments, Internal validity, Procedural concerns, Psychological game theory.

**JEL Classification:** C70, C93, D63, I38, O12.

[*] We thank Angus Deaton, Pramila Krishnan, Claudia Senik, and seminar participants at the Cambridge, Konstanz, Helsinki, and Paris School of Economics for useful comments.

[†] Department of Economics and CRED, University of Namur, Rempart de la Vierge 8, 5000 Namur, Belgium. E-mail: gani.aldashev@fundp.ac.be

[‡] ECARES, Université libre de Bruxelles, Avenue F D Roosevelt 50, CP114 1050 Brussels, Belgium, ECORE, CEPR, and CESifo. E-mail: gkirchst@ulb.ac.be

[§] Department of Economics, University of Copenhagen, Øster Farimagsgade 5, 1353 Copenhagen, Denmark. E-mail: alexander.sebald@econ.ku.dk

# 1 Introduction

Randomized 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 labor 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 health outcomes of children in Mexico (Schultz (2004), Gertler (2004)), of vouchers for private schooling on school completion rates in Colombia (Angrist et al. (2002), Angrist et al. (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 (Michalopoulos et al. (2005), Card and Robins (2005), Card and Hyslop (2005)), of saving incentives on the saving decisions of low- and middle-income families in the United States (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 program 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 behavior 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 program. Based on the experimental results, the program 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 external validity. Factors that have been shown to potentially influence the external validity of RCTs are among others the randomization bias (Heckman 1991), Hawthorn and John Henry effect (see e.g. Levitt and List (2007) and Duflo, Glennerster and Kremer (2008)) and the placebo effect (Malani (2006)).

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

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

> "*Changes in behavior among the treatment group are called Hawthorne effects, while changes in behavior 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 behavior for the duration of the experiment (for example, working harder to make it a success). The comparison group may feel offended to be a comparison group and react by also altering their behavior (for example, teachers in the comparison group for an evaluation may 'compete' with the treatment teachers or, on the contrary, decide to slack off).*"

Building on this intuition and building on the recent literature on belief-dependent preferences (see e.g. Geanakoplos et al. (1989), Battigalli and Dufwenberg (2009)), we theoretically analyze such feelings of encouragement (of the treatment group) and resentful demoralization (of the control group) as described by Duflo, Glennerster and Kremer (2008) and show that these are two facets of the same behavioral root - namely people's propensity to act reciprocally. More specifically, to theoretically analyze the impact of encouragement and resentful demoralization 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) in which agents are motivated by belief-dependent reciprocity.[3] That is,

[2] As evidence on this point for the Hawthorn effect Levitt and List (2011), for example, write: "The Merriam-Webster dictionary offers a very different definition for a Hawthorne effect than the one cited in the OED: "the stimulation to output or accomplishment that results from the mere fact of being under observation." Along the lines of this second definition, is the use of so-called 'Hawthorne control groups' in randomized trials in addition to the standard treatment and control groups." (Levitt and List (2011, p. 227))

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

agents 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 analyze feelings of encouragement and resentful demoralization in RCTs, but also delivers intriguing insights regarding their potential character and importance. We show, for example that it is not primarily the fact that people are 'under scrutiny' or 'under observation' in RCTs that drives these biases, but that the assignment procedure used to allocated people into control and treatment group crucially determines their size.

In line with Duflo, Glennerster and Kremer (2008)'s loose definition of the Hawthorn and John Henry effect we find that a subject motivated by reciprocity not assigned to the treatment group (while other similar agents are) feels discouraged and provides less effort than without the existence 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), she 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.

Interestingly, the size and potential sign 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) randomization procedure, the amount of resentful demoralization (encouragement) is particularly high. The estimate of the effect of a general introduction of the treatment under this type of randomization procedure is unambiguously biased upwards. On the other hand, if the experimenter uses an explicit and credible randomization mechanism, the impact of demoralization and encouragement is lower. Hence, the problem of the upward bias in the estimate is reduced. However, an unbiased and transparent randomization procedure might also lead to a negative bias, i.e. an under-estimation of the true estimate. That is, 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

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

introduction of the treatment. Notwithstanding this finding, we show that unbiased and credible randomization procedures that allocate people into treatment and control group minimize the bias in the estimate of the true effect relative to biased and less transparent assignment mechanisms.

These behavioral effects are not just hypothetical. A change in the behavior of the control group is well-known in psychology under the heading 'resentful demoralization'.[5] A good example where this demoralization effect played a key role is the Birmingham Homeless Project (Schumacher et al. (1994)), aimed at homeless drug-takers in Birmingham, Alabama. The randomly assigned subjects of the treatment group received more frequent and therapeutically superior treatment, as compared to those in the control group. Schumacher et al. (1994, p.42) note that "11 percent increase in cocaine relapse rates for usual care clients [i.e. the control group] was revealed". They conclude, "demoralization 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 demoralization, there exists a potential for an overestimation of treatment effects of the enhanced care program" (Schumacher et al. (1994, p.43-44)).

Another example is the Baltimore Options Program (Friedlander et al. (1985)), which was designed to increase the human capital and, hence, the employment possibilities of unemployed young welfare recipients in the Baltimore Country. Half of the potential recipients were randomly assigned to the treatment group and half to the control group. The treatment group individuals received tutoring and job search training for one year. The control group members, aware of not having received the (desirable) treatment, became discouraged and thus performed worse in the outcome measure than they would have performed if the treatment group did not exist. This clearly leads to an overestimation of the effectiveness of the program. In fact, researchers found that the earnings of the treatment group increased by 16 percent, but that the overall welfare claims of program participants did not decrease. This implies that some of the control-group individuals that would have normally moved out of welfare stayed longer on welfare because of the experiment.

[5] This was first described in detail by Cook and Campbell in their seminal 1979 book, where they list resentful demoralization among potential threats to the internal validity of experiments in social sciences. See Onghena (2009) for a short survey.

The treatment group, on the other hand, might be encouraged to perform particularly well (the Hawthorn effect as described by ...). Take a young unemployed chosen to participate in the Baltimore Options Program. The very fact of being picked among all the other young unemployed (and this, *without* explaining to him that choosing him was the outcome of a randomization procedure) might encourage him, and this encouragement can induce him to intensify his job search. Empirically, this effect is hard to distinguish from the effort increase caused by the better job market perspectives induced by the enhanced skills from the training program. Nonetheless, the existence of such an effect seems plausible in numerous experimental settings and hence worth analyzing its possible impact on the estimates obtained.

This paper contributes to the small but growing economic literature theoretically analyzing the behavior of subjects in RCTs.[6] The closest papers to ours are by 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 thus 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 coming from 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 her *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. In this setting, more optimistic patients respond stronger to treatment than the less optimistic ones, and 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. The paper then proposes a solution for this problem which consists of designing the experiment with two (or more) treatment

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

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 corresponds to the detection 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 but the outcome depends on the (costly) effort and returns multiplicatively. The estimate of returns obtained from such an experiment would be imprecise because of the unobservable heterogeneity of expectations and thus of effort exerted by subjects. The solution that the authors propose 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 demoralization and encouragement effect created by the assignment procedure, i.e. the Hawthorn and John Henry effect, - an issue not analyzed in the above literature.

In the next section we present a simple model of policy experiments that takes the demoralization and encouragement into account. Section 3 derives formally the biases connected to the different randomized assignment procedures. The final section discusses the implication of our results for the design of RCTs and concludes.

## 2 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, or 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). Denote the overall population of agents by $N$. $n$ of these agents are subject to the treatment, and $q = \frac{n}{N} \in [0, 1]$ denotes the fraction of agents in the treatment group.

To concentrate on the impact of the randomized 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.

(ii) The experimenter can choose an explicit randomization procedure observable to the agents, such that each agent has the same probability $q$ of receiving the treatment. This also models a closed-door random assignment procedure, when the subjects do not doubt the randomness of the assignment.

Since we are interested in the impact of the assignment procedure, we will not analyze the experimenter's equilibrium choice as if she 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 equivalent, 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 randomization 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 $e \in [0, 1]$.

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 $t = \frac{1}{2}$ for the other agents. Thus, the treatment makes it easier for participants to be successful. We use the variable $t \in \{\frac{1}{2}, 1\}$ to denote also whether an agent is in the control group ($t = \frac{1}{2}$) or 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

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

As already mentioned, we do not analyze the experimenter's equilibrium choice as if she 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 would know that the effectiveness of the program is tested and that the experimental results determine the long-run feasibility and shape of the program, the agents' long-term strategic interests would jeopardize the validity of the experimental results. To give the randomized experiments the best shot, we abstract from such effects by assuming that the agents, unaware of the experimental character of the program, consider the overall success, denoted by $\pi_x$, as the goal of the experimenter.[7] It depends on the effort levels chosen by the agents (which, in turn, depends on the assignment procedure), and on the group sizes:

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

We assume that the agents are motivated by their individual success: unemployed want to find a job, the 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:

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

[7] The exact form of the experimenter's goal as perceived by the agent is not important for our results. Any goal function would lead to allocation biases, as long as the agents believe that the experimenter cares about the agent's success.

However, as we argue above, agents do not only care about their material payoffs, but also about the way they are treated. If an agent feels treated badly, she resents the experimenter, feels discouraged, and hence, is less willing to provide effort. On the other hand, if the agent feels treated particularly well, she might feel encouraged, may want the program 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 she 'thinks' that the experimenter 'intends' to give her relative to a 'neutral' material payoff.

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

$$\overline{e}^{t,p} = (\overline{e}^{t,p}(1, d), \overline{e}^{t,p}(\frac{1}{2}, d), \overline{e}^{t,p}(1, r), \overline{e}^{t,p}(\frac{1}{2}, r)).$$

Furthermore, let $\overline{\overline{e}}^{t,p}(t', p')$ denote the *second-order belief* of a $(t, p)$-agent about the experimenter's belief concerning the effort choice of a $(t', p')$. The second-order beliefs of a $(t, p)$-agent are summarized by

$$\overline{\overline{e}}^{t,p} = (\overline{\overline{e}}^{t,p}(1, d), \overline{\overline{e}}^{t,p}(\frac{1}{2}, d), \overline{\overline{e}}^{t,p}(1, r), \overline{\overline{e}}^{t,p}(\frac{1}{2}, r)).$$

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

$$\pi_x(e(t, p), \overline{e}^{t,p}) = \begin{cases} e(1, p) + (n-1) \cdot \overline{e}^{1,p}(1, p) + (N-n) \cdot \frac{1}{2} \cdot \overline{e}^{1,p}(\frac{1}{2}, p) & \text{if } t = 1 \\ \frac{1}{2} \cdot e(\frac{1}{2}, p) + n \cdot \overline{e}^{\frac{1}{2},p}(1, p) + (N-n-1) \cdot \frac{1}{2} \cdot \overline{e}^{\frac{1}{2},p}(\frac{1}{2}, p) & \text{if } t = \frac{1}{2} \end{cases} \tag{4}$$

Note that $\pi_x(e(t, p), \overline{e}^{t,p})$ 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 $\frac{\partial \pi_x(e(t,p),\overline{e}^{t,p})}{\partial e(t,p)} = t$.

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

$$\pi(\overline{\overline{e}}^{t,p}) = \begin{cases} q \cdot (\overline{\overline{e}}^{t,r}(1,r) - \overline{\overline{e}}^{t,r}(1,r)^2) + (1-q) \cdot (\frac{1}{2} \cdot \overline{\overline{e}}^{t,r}(\frac{1}{2},r) - \overline{\overline{e}}^{t,r}(\frac{1}{2},r)^2) & \text{if } p = r \\ t \cdot \overline{\overline{e}}^{t,d}(t,d) - \overline{\overline{e}}^{t,d}(t,d)^2 & \text{if } p = d \end{cases} \tag{5}$$

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

We also have to specify the 'neutral' payoff $\widehat{\pi}$ at which the agent regards the principal's choice of assignment procedure as being materially neutral, i.e. neither favoring 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 her $\widehat{\pi}$, she is neither discouraged nor encouraged, and hence she simply maximizes her material payoff.

Note that the expected material payoff of an agent is maximized when she is directly assigned to the treatment group. It is minimized when the agent is directly assigned to the control group. Therefore, we assume that $\widehat{\pi}$ 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 $\lambda$ and $1-\lambda$, respectively,

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

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

with $\lambda \in [0,1]$:

$$\widehat{\pi}(\overline{\overline{e}}^{t,p}) = \lambda \cdot (\overline{\overline{e}}^{t,p}(1,d) - \overline{\overline{e}}^{t,p}(1,d)^2) + (1-\lambda) \cdot (\frac{1}{2} \cdot \overline{\overline{e}}^{t,p}(\frac{1}{2},d) - \overline{\overline{e}}^{t,p}(\frac{1}{2},d)^2), \tag{6}$$

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

The weight $\lambda$ depends on the fraction of agents that are subject to the treatment, i.e. $q$. Whenever a randomized 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 program might think that they deserve 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 demoralization and encouragement, we assume that the higher the payoff $\pi(\overline{\overline{e}}^{t,p})$ that the agent believes the experimenter intends to give her (as compared to the neutral payoff $\widehat{\pi}(\overline{\overline{e}}^{t,p})$), the more encouraged and the less resentful she is. Denoting by $v(\pi_x(e(t,p), \overline{e}^{t,p}), \pi(\overline{\overline{e}}^{t,p}), \widehat{\pi}(\overline{\overline{e}}^{t,p}))$ the psychological payoff in the agent's utility derived from demoralization and encouragement, a simple way to capture these motives is by assuming that

$$\frac{\partial v(\pi_x(e(t,p), \overline{e}^{t,p}), \pi(\overline{\overline{e}}^{t,p}), \widehat{\pi}(\overline{\overline{e}}^{t,p}))}{\partial \pi_x} = \pi(\overline{\overline{e}}^{t,p}) - \widehat{\pi}(\overline{\overline{e}}^{t,p}). \tag{7}$$

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

Summarizing, the belief-dependent utility of a reciprocal $(t,p)$-agent is the sum of

[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).

the material and the psychological payoffs:

$$u^{t,p}(e(t,p),\overline{e}^{t,p},\overline{\overline{e}}^{t,p}) = t\cdot e(t,p) - e(t,p)^2 + v(\pi_x(e(t,p),\overline{e}^{t,p}),\pi(\overline{\overline{e}}^{t,p}),\widehat{\pi}(\overline{\overline{e}}^{t,p})). \qquad (8)$$

This closes the description of our stylized randomized control trial with reciprocal agents. Next we analyze the impact of the procedure on the agents' behavior.

# 3 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 maximizes her utility for first- 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 an equilibrium in pure strategies. The equilibrium effort levels are in the interior, i.e.* $0 < e^*(t,p) < 1$ *for all* $t,p$.

**Proof:** See Appendix.

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 randomization procedure.

**Proposition 2** *For any fraction of agents in the treatment group* $q \in (0,1)$ :

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

**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 randomization) than when it is done using a

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

public randomization 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 shows that the randomization procedure has an impact on the observed effectiveness of the treatment. On the one hand, agents feel encouraged if they think that they are deliberately chosen to get the treatment. On the other hand, agents are more discouraged when they feel deliberately assigned into the control group. This result holds for any fraction of people that are assigned into the treatment group $q \in (0, 1)$.

The previous proposition shows that randomization procedures have an impact on the behavior of agents in policy experiments. The key question then is: which procedure provides a correct (i.e. internally valid) 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 program 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 randomization procedures.

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

**Proof:** See Appendix

Proposition 6 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 she were motivated only by her material payoff. Proposition 6 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 randomization procedure *always* leads to an overestimation of the impact of the treatment, as the following proposition shows.

**Proposition 4** *For any fraction of agents in the treatment group $q \in (0, 1)$,*

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

**Proof:** See Appendix.

Under a private randomization assignment, the effort level of the control group is always smaller than the effort level realized when the entire population does not receive the treatment. The effort level of the treatment group is always larger than the one realized 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 randomization is biased upwards. A policy-maker scaling up the program on the basis of such an RCT faces the risk of introducing a non-effective program to the entire population.

One might hope that with an explicit and credible randomization 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 does not need to be the case.

**Proposition 5**

*i) For any $\lambda \in (0, 1)$, there exists at most one $q$ such that $e^*(1, r) - e^*(\frac{1}{2}, r) = e^*_1 - e^*_0 = \frac{1}{4}$.*

*ii) If $\lambda = q \in (0, 1)$, $e^*(1, r) - e^*(\frac{1}{2}, r) \neq \frac{1}{4}$.*

**Proof:** See Appendix.

Explicit randomization 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 resulting neutral payoff should equal the expected material payoff of an agent subject to explicit randomization. Hence, even under a public randomization the experimental results do not reflect the true benefits of a general introduction of the treatment. This is in particular true for the natural case of $\lambda = q$ when agents have a 'rational' perception of how much they deserve the treatment.

Importantly, however, while explicit randomization does not completely solve the problem of a biased estimation of the true impact of the treatment, the following result shows that it certainly minimizes 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:

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

Using this variable, we can state the following result:

**Proposition 6** *If $\lambda = q \in (0,1)$, then $\left|b^d\right| > |b^r|$.*

**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 randomization procedure relative to a direct appointment mechanism.

# 4 Conclusion

In this paper we have analyzed feelings of encouragement and resentful demoralization, two expressions of the Hawthorn and John Henry effect, their common behavioral root and their impact on the external validity of policy experiments. We show, if agents are prone to demoralization and encouragement, the way in which experimenters assign them into the treatment and control groups influences their behavior. Thus, the size of the treatment effect depends on the assignment procedure. If agents are directly assigned into the treatment and control group, or if agents believe that they are directly assigned, the experimentally observed treatment effect is always larger than the effect of a general introduction of the treatment. Although not completely resolving it, this assignment procedure bias is always smaller for a credible (explicit) randomization procedure.

This analysis concentrates on the effects of reciprocity, and hence, demoralization and encouragement. There are other belief-dependent motives like guilt (see e.g.

Charness and Dufwenberg (2006), and Battigalli and Dufwenberg (2007)), or disappointment (see e.g. Ruffle (1999)) that have been found to impact agents' behavior. Exploring the impact of these effects on the external validity of RCTs is left to future research.

# 5 Appendix

## Proof of proposition 1

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

$$\begin{aligned} \frac{\partial u^{t,p}(e(t,p),\overline{e}^{t,p},\overline{\overline{e}}^{t,p})}{\partial e(t,p)} &= t(1+v^{t,p}_{\pi x}) - 2e(t,p), && (10) \\ \frac{\partial^2 u^{t,p}(e(t,p),\overline{e}^{t,p},\overline{\overline{e}}^{t,p})}{\partial e(t,p)^2} &= \frac{\partial^2 v(\pi_x(e(t,p),\overline{e}^{t,p}),\pi(\overline{\overline{e}}^{t,p}),\widehat{\pi}(\overline{\overline{e}}^{t,p}))}{(\partial \pi_x)^2} t^2 - 2. && (11) \end{aligned}$$

Since $\frac{\partial^2 v(\pi_x(e(t,p),\overline{e}^{t,p}),\pi(\overline{\overline{e}}^{t,p}),\widehat{\pi}(\overline{\overline{e}}^{t,p}))}{(\partial \pi_x)^2} = 0$,

$$\frac{\partial^2 u^{t,p}(e(t,p),\overline{e}^{t,p},\overline{\overline{e}}^{t,p})}{\partial e(t,p)^2} < 0 \quad \text{for all } t,p. \tag{12}$$

Because $|v^{t,p}_{\pi x}| \leq \frac{3}{4}$, it is easy to check that

$$\begin{aligned} \left.\frac{\partial u^{t,p}(e(t,p),\overline{e}^{t,p},\overline{\overline{e}}^{t,p})}{\partial e(t,p)}\right|_{e(t,p)=0} &> 0 \quad \text{for all } t,p, \\ \left.\frac{\partial u^{t,p}(e(t,p),\overline{e}^{t,p},\overline{\overline{e}}^{t,p})}{\partial e(t,p)}\right|_{e(t,p)=1} &< 0 \quad \text{for all } t,p. \end{aligned} \tag{13}$$

Because of (12) and (13), each of the equations

$$\frac{\partial u^{t,p}(e(t,p),\overline{e}^{t,p},\overline{\overline{e}}^{t,p})}{\partial e(t,p)} = 0 \tag{14}$$

has a unique interior solution for each $t,p$ for any first- and second-order belief $\overline{e}^{t,p}, \overline{\overline{e}}^{t,p}$. These solutions characterize the optimal effort choices of all types of agents for given first- and second-order beliefs. In equilibrium, the beliefs of first- and second-order have to be the same, i.e. $\overline{e}^{t,p} = \overline{\overline{e}}^{t,p}$ for all $t, p$. The solution of (14) can be rewritten as a function

$$e^{t,p}_{opt} : [0,1]^4 \rightarrow [0,1]^4,$$

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

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

The effort levels characterized by this fixed point maximize the agents' utilities for first- and second-order beliefs which coincide with the utility maximizing effort levels, i.e. for correct beliefs. Hence, $e^*$ fulfills the conditions for an equilibrium.■

## Proof of proposition 2

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

$$\begin{aligned} 1 - 2e(1,d) + v^{1,d}_{\pi x} &= 0, & (15)\\ \frac{1}{2} - 2e(\frac{1}{2},d) + v^{\frac{1}{2},d}_{\pi x}\frac{1}{2} &= 0, & (16)\\ 1 - 2e(1,r) + v^{1,r}_{\pi x} &= 0, & (17)\\ \frac{1}{2} - 2e(\frac{1}{2},r) + v^{\frac{1}{2},r}_{\pi x}\frac{1}{2} &= 0. & (18) \end{aligned}$$

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

To prove the proposition, we first show that $e^*(1,r) > e^*(\frac{1}{2},r)$. Since in equilibrium $\overline{\overline{e}}^{\frac{1}{2},r}(t',p') = \overline{\overline{e}}^{1,r}(t',p') = e(t',p')$, $\overline{\overline{\pi}}^{1,r}_a(\overline{\overline{e}}^{1,r}) = \overline{\overline{\pi}}^{\frac{1}{2},r}_a(\overline{\overline{e}}^{\frac{1}{2},r})$. Because of this equality, $v^{1,r}_{\pi x} = v^{\frac{1}{2},r}_{\pi x}$. Using this and comparing the FOCs (17) and (18) reveal that $e^*(1,r) = 2e^*(\frac{1}{2},r) > e^*(\frac{1}{2},r)$.

Second, we prove that

$$e^*(1,r) - e^*(1,r)^2 > \frac{1}{2}e^*(\frac{1}{2},r) - e^*(\frac{1}{2},r)^2. \tag{19}$$

Inserting $e^*(1,r) = 2e^*(\frac{1}{2},r)$ and rearranging terms, (19) 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 equations (5), (7) and (19) it is true that

$$\begin{array}{rcl} v_{\pi x}^{1,d} - v_{\pi x}^{1,r} & = & e(1,d) - e(1,d)^2 - q(e(1,r) - e(1,r)^2) - (1-q)(\frac{1}{2}e(\frac{1}{2},r) - e(\frac{1}{2},r)^2) \\ & > & e(1,d) - e(1,d)^2 - e(1,r) + e(1,r)^2. \end{array}$$

Comparing (15) to (17), one sees that

$$v_{\pi x}^{1,d} - v_{\pi x}^{1,r} = 2(e(1,d) - e(1,r)), \tag{20}$$

implying that

$$e(1,d) - e(1,r) > -e(1,d)^2 + e(1,r)^2. \tag{21}$$

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

Finally, it remains to show that $e^*(\frac{1}{2},r) > e^*(\frac{1}{2},d)$. Because of equations (5), (7) and (19), it holds that

$$\begin{array}{rcl} v_{\pi x}^{\frac{1}{2},r} - v_{\pi x}^{\frac{1}{2},d} & = & q(e(1,r) - e(1,r)^2) + (1-q)(\frac{1}{2}e(\frac{1}{2},r) - e(\frac{1}{2},r)^2) - \frac{1}{2}e(\frac{1}{2},d) + e(\frac{1}{2},d)^2) \\ & > & \frac{1}{2}(e(\frac{1}{2},r) - e(\frac{1}{2},d)) - e(\frac{1}{2},r)^2 + e(\frac{1}{2},d)^2). \end{array}$$

Comparing (16) to (18), one gets

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

implying that

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

However, this condition can only hold for $e^*(\frac{1}{2},r) > e^*(\frac{1}{2},d)$■

## Proof of proposition 6

i) $q=1$ implies that $\lambda=1$. Therefore, $\pi(\overline{\overline{e}}^{1,d})=\widehat{\pi}(\overline{\overline{e}}^{1,d})$ and $v_{\pi x}^{1,d}=0$. From (15) follows that $e^*(1,d)=\frac{1}{2}$. Since the beliefs have to be correct in equilibrium, we get that $\widehat{\pi}(\overline{\overline{e}}^{1,r})=\frac{1}{4}$. By substituting into (17) we get

$$1-2e(1,r)+(e(1,r)-e(1,r)^2-\frac{1}{4})=0, \tag{22}$$

given that the beliefs have to be correct. The unique solution to (22) is $e^*(1,r)=\frac{1}{2}$.

ii) $q=0$ implies that $\lambda=0$. Therefore, $\pi(\overline{\overline{e}}^{\frac{1}{2},d})=\widehat{\pi}(\overline{\overline{e}}^{\frac{1}{2},d})$ and $v_{\pi x}^{1,d}=0$. From (16) follows that $e^*(1,d)=\frac{1}{4}$. Since the beliefs have to be correct in equilibrium, we get that $\widehat{\pi}(\overline{\overline{e}}^{1,r})=\frac{1}{16}$. By substituting into (18) we get

$$\frac{1}{2}-2e(\frac{1}{2},r)+\frac{1}{2}(\frac{1}{2}e(\frac{1}{2},r)-e(\frac{1}{2},r)^2-\frac{1}{16})=0, \tag{23}$$

given that the beliefs have to be correct. The unique solution to (23) is $e^*(\frac{1}{2},r)=\frac{1}{4}$. ■

## 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

$$\begin{aligned} v_{\pi x}^{1,d} &= (1-\lambda)(e(1,d)-e(1,d)^2-\frac{1}{2}e(\frac{1}{2},d)+e(\frac{1}{2},d)^2), \\ v_{\pi x}^{\frac{1}{2},d} &= -\lambda(e(1,d)-e(1,d)^2-\frac{1}{2}e(\frac{1}{2},d)+e(\frac{1}{2},d)^2) \end{aligned} \tag{24}$$

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 $\frac{1}{2}$ and $\frac{1}{4}$, respectively (see FOCs (15) and (16)). 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 (15) and (16) implies that $e(1,d)<\frac{1}{2}$ and $e(\frac{1}{2},d)>\frac{1}{4}$. Since $e(1,d)>e(\frac{1}{2},d)$, this implies that $e(t,d)\in(\frac{1}{4},\frac{1}{2})$ for $t=1,\frac{1}{2}$.

Because of (24) and $v_{\pi x}^{1,d} < 0 < v_{\pi x}^{\frac{1}{2},d}$,

$$-e(1,d) + e(1,d)^2 + \frac{1}{2}e(\frac{1}{2},d) - e(\frac{1}{2},d)^2 = -v_{\pi x}^{1,d} + v_{\pi x}^{\frac{1}{2},d} > 0 \tag{25}$$

For $e(t,d) \in (\frac{1}{4},\frac{1}{2})$ the left-hand side of (25) is decreasing in $e(1,d)$ and $e(\frac{1}{2},d)$. However, even for the limit case of $e(1,d) = e(\frac{1}{2},d) = \frac{1}{4}$ the left hand side of (25) is $-\frac{1}{8}$. Hence (25) 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 (24) also implies that $e^*(1,d) - e^*(1,d)^2 > \frac{1}{2}e^*(\frac{1}{2},d) - e^*(\frac{1}{2},d)^2$ - the material payoff from getting a treatment is larger than from not getting a treatment, if the selection is done directly.

Recall that $v_{\pi x}^{t,p} \in [-\frac{3}{4},\frac{3}{4}]$. Hence, $v_{\pi x}^{1,d} \in (0,\frac{3}{4}]$ and $v_{\pi x}^{\frac{1}{2},d} \in [-\frac{3}{4},0)$. Using this and the FOCs (15) and (16) one immediately gets that $e^*(1,d) \in (\frac{1}{2},\frac{7}{8}]$ and that $e^*(\frac{1}{2},d) \in [\frac{1}{16},\frac{1}{4})$. ■

## Proof of proposition 5

i) Subtracting (18) from (17) reveals that $v_{\pi x}^{1,r} - \frac{1}{2}v_{\pi x}^{\frac{1}{2},r} = 0$, whenever in equilibrium $e(1,r) - e(\frac{1}{2},r) = \frac{1}{4}$. Since $v_{\pi x}^{1,r} = v_{\pi x}^{\frac{1}{2},r}$, this can only hold for $v_{\pi x}^{1,r} = v_{\pi x}^{\frac{1}{2},r} = 0$. Hence, $e(1,r) = \frac{1}{2}$, $e(\frac{1}{2},r) = \frac{1}{4}$ in equilibrium if the difference in equilibrium effort is $\frac{1}{4}$.

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

$$\widehat{\pi} = \frac{3q+1}{16}. \tag{26}$$

Using the definition of $\widehat{\pi}$, (26), and again the fact that the equilibrium beliefs are correct, we get

$$\frac{3q+1}{16} = \lambda\pi(1,d) + (1-\lambda)\pi(\frac{1}{2},d). \tag{27}$$

If in equilibrium $e(1,r) - e(\frac{1}{2},r) = \frac{1}{4}$, then the equation (27) has to hold. Recall that $\pi(1,d)$ and $\pi(\frac{1}{2},d)$ are determined by the joint solution of the FOCs (15) and (16). Since $v_{\pi x}^{t,d}$ is independent of $q$, $\pi(1,d)$ and $\pi(\frac{1}{2},d)$ do not depend on $q$. Hence the right-hand side of (27) 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 $e_1^* - e_0^* = \frac{1}{4}$.

ii) Inserting (26) into (7) and (15) leads to

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

By solving this equation one gets

$$e(1,d) = \frac{-2+\sqrt{19-3q}}{4}. \tag{28}$$

Inserting 26) into 7) and 16) leads to

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

By solving this equation one gets

$$e(\frac{1}{2},d) = \frac{-7+\sqrt{64-3q}}{4} \tag{29}$$

Given that $\lambda = q$ and because of (29) and (28), (27) becomes

$$\begin{aligned} \frac{3q+1}{16} &= q\left(\frac{-2+\sqrt{19-3q)}}{4} - \left(\frac{-2+\sqrt{19-3q)}}{4}\right)^2\right) \\ &\quad +(1-q)\left(\frac{1}{2}\frac{-7+\sqrt{64-3q}}{4} - \left(\frac{-7+\sqrt{64-3q}}{4}\right)^2\right), \end{aligned} \tag{30}$$

leading to

$$0 = 96q + 8q\sqrt{19-3q} - 16q\sqrt{64-3q} + 16\sqrt{64-3q} - 128. \tag{31}$$

For any $q \in (0,1)$, the right-hand side of (31) is strictly larger than zero. This equation holds only for the limit cases $q = 1$ and $q = 0$.

## Proof of proposition 6

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

a) $b^r \geq 0$. In this case Proposition 4 implies that $\left|b^d\right| > \left|b^d\right|$.

b) $b^r < 0$.

From the definition (9), the first order conditions (15)-(18) and taking into account that $v_{\pi x}^{1,r} = v_{\pi x}^{\frac{1}{2},r} = v_{\pi x}^{r}$ we get

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

Because $b^r < 0$, the proposition holds if

$$b^d + b^r = \frac{2v_{\pi x}^{1,d} - v_{\pi x}^{\frac{1}{2},d} + v_{\pi x}^{r}}{4} > 0. \tag{32}$$

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

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

Inserting this, condition (32), the condition for the proposition to hold becomes

$$\begin{aligned} 2(1-\lambda)\left(\pi(1, e(1,d)) - \pi(\frac{1}{2}, e(\frac{1}{2}, d))\right) + \lambda\pi(1, e(1,r)) \\ + (1-\lambda)\pi(\frac{1}{2}, e(\frac{1}{2}, r)) - \pi(\frac{1}{2}, e(\frac{1}{2}, d)) > 0 \end{aligned} \tag{33}$$

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 33 holds. ■

# 6 References

1. 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 Heckman, J., and Leamer, E. (eds.), Handbook of Econometrics, vol. 6, Elsevier, Amsterdam.

2. 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, 92, 1535-1558.

3. 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, 96, 847-862.

4. 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.

5. Battigalli, P., and Dufwenberg, M. (2007), *Guilt in Games*, American Economic Review, Papers and Proceedings, 97, 170–176.

6. Battigalli, P. and Dufwenberg, M. (2009), *Dynamic Psychological Games*, Journal of Economic Theory, 144, 1-35.

7. 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, 125, 113-139.

8. Card, D., and Hyslop, D. (2005), *Estimating the Effects of a Time-Limited Earnings Subsidy for Welfare-Leavers*, Econometrica, 73, 1723-1770.

9. Charness, G., and Dufwenberg, M. (2006), *Promises and Partnership*, Econometrica, 74, 1579-1601.

10. Chassang, S., Padró-i-Miquel, G., and Snowberg, E. (2012), *Selective Trials: A Principal-Agent Approach to Randomized Controlled Experiments*, American Economic Review, 102, 1279-1309.

11. Cook, T., and Campbell, D. (1979), *Quasi-Experimentation: Design and Analysis Issues for Field Settings*, Houghton Mifflin, Boston, MA.

12. De Mel, S., McKenzie, D., and Woodruff, C. (2008), *Returns to Capital in Microenterprises: Evidence from a Field Experiment*, Quarterly Journal of Economics, 123, 1329-1372.

13. Duflo, E. (2004), *Scaling Up and Evaluation*, Annual World Bank Conference on Development Economics, The World Bank, Washington, DC.

14. 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, 121, 1311-1346.

15. Duflo, E., Glennerster, R. and Kremer, M. (2008), *Using Randomization in Development Economics Research: A Toolkit.* In Handbook of Development Economics, Vol. 4, ed. T. Paul Schultz and John A. Strauss, 3895-3962. Amsterdam: Elsevier.

16. Dufwenberg, M., and Kirchsteiger, G. (2004), *A Theory of Sequential Reciprocity*, Games and Economic Behavior, 47, 268-298.

17. Ferraz, C., and Finan, F. (2008), *Exposing Corrupt Politicians: The Effects of Brazil's Publicly Released Audits on Electoral Outcomes*, Quarterly Journal of Economics, 123, 703-745.

18. Friedlander, D., Hoetz, G., Long, D., and Quint, J. (1985), *Maryland: Final Report on the Employment Initiatives Evaluation*, MDRC, New York, NY.

19. Geanakoplos, J., Pearce, D., and Stacchetti, E. (1989), *Psychological games and sequential rationality*, Games and Economic Behavior, 1, 60–79.

20. Gertler, P. (2004), *Do conditional cash transfers improve child health? Evidence from PROGRESA's controlled randomized experiment*, American Economic Review, 94, 336-341.

21. Heckman, J. (1991) *Randomization and Social Policy Evaluation.* National Bureau of Economic Research Technical Working Paper 107.

22. Heckman, J., and Vytlacil, E. (2007a), *Econometric Evaluation of Social Programs, Part I: Causal Models, Structural Models and Econometric Policy Evaluation*, in Heckman, J., and Leamer, E. (eds.), Handbook of Econometrics, vol. 6, Elsevier, Amsterdam.

23. Heckman, J., and Vytlacil, E. (2007b), *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 Heckman, J., and Leamer, E. (eds.), Handbook of Econometrics, vol. 6, Elsevier, Amsterdam.

24. Hoorens, V. (1993), *Self-enhancement and superiority biases in social comparison*, European review of social psychology, 4(1), 113-139.

25. Imbens, G., and Wooldridge, J. (2009), *Recent Developments in the Econometrics of Program Evaluation*, Journal of Economic Literature, 47, 5-86.

26. Levitt, S., and List, J. (2011), *Was There Really a Hawthorne Effect at the Hawthorse Plant? An Analysis of the Original Illumination Experiments*, American Economic Journal: Applied Economics, 3, 224-238.

27. Malani, A. (2006), *Identifying Placebo Effects with Data from Clinical Trials*, Journal of Political Economy, 114, 236-256.

28. 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, 89, 5-29.

29. Onghena, S. (2009), *Resentful demoralization*, in Everitt B., Howel D. (eds.), Encyclopedia of statistics in behavioral science, vol. 4, Wiley, Chichester, UK.

30. Philipson, T., and Hedges, L. (1998), *Subject Evaluation in Social Experiments*, Econometrica, 66, 381-408.

31. Rabin, M. (1993), *Incorporating Fairness into Game Theory and Economics*, American Economic Review, 83, 1281-1302.

32. Ruffle, B. (1999), *Gift giving with emotions*, Journal of Economic Behavior & Organization, 39, 399–420.

33. Schultz, T.P. (2004), *School subsidies for the poor: evaluating the Mexican Progresa poverty program*, Journal of Development Economics, 74, 199-250.

34. 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 Conrad, K. (ed.), Critically Evaluating the Role of Experiments, Jossey-Bass, San Francisco, CA.

35. Sebald, A. (2010), *Attribution and Reciprocity*, Games and Economic Behavior, 68, 339-352.