NBER WORKING PAPER SERIES

TOWARD AN UNDERSTANDING OF THE WELFARE EFFECTS OF NUDGES: EVIDENCE FROM A FIELD EXPERIMENT IN UGANDA

Erwin Bulte
John A. List
Daan Van Soest

Working Paper 26286
http://www.nber.org/papers/w26286

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
September 2019

The authors thank John Sseruyange for his excellent research assistance when implementing the experiments in the field. List thanks the Sloan Foundation for financial support. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Toward an Understanding of the Welfare Effects of Nudges: Evidence from a Field Experiment in Uganda

Erwin Bulte, John A. List, and Daan Van Soest

NBER Working Paper No. 26286
September 2019
JEL No. C93,D03,J01

ABSTRACT

Social scientists have recently explored how framing of gains and losses affects productivity. We conducted a field experiment in peri-urban Uganda, and compare output levels across 1000 workers over isomorphic tasks and incentives, framed as either losses or gains. We find that loss aversion can be leveraged to increase the productivity of labor. The estimated welfare costs of using the loss contract are quite modest – perhaps because the loss contract is viewed as a (soft) commitment device.

Erwin Bulte
Wageningen University
Department of Social Sciences
Hollandseweg 1
6706 Kn WAGENINGEN
Netherlands
erwin.bulte@wur.nl

Daan Van Soest
Tilburg University
d.p.vansoest@uvt.nl

John A. List
Department of Economics
University of Chicago
1126 East 59th
Chicago, IL 60637
and NBER
jlist@uchicago.edu

1. Introduction

Constructing incentive schemes to motivate others to take a particular course of action is a key challenge across societies. Whether at home, at school, in the office, or in public spaces, we are surrounded by incentives created to induce a particular set of behaviours. For their part, economists have produced a rich assortment of models to help understand situations in which the various elements of incentive contracts should enforce productive behaviours, and when they might exacerbate misconduct. As an important complement to standard theory, the field of behavioural economics is evolving rapidly to explain departures from rational models. Early work focused on lab experiments documenting deviations from "rational behaviour" – *Homo economicus* style – and subsequently attempted to capture "behavioural anomalies" in formal models of preferences and beliefs.

In recent years the profession has increasingly combined psychology and economics in analyses of people interacting in organizations and markets. This line of work considers both efficiency and distributional issues, and seeks to explore how firms and governments can advance their aims by taking insights from behavioural economics more seriously when constructing incentive contracts. While some analyses focus on so-called exploitative contracts, where firms seek to take advantage of behavioural "mistakes" (e.g., DellaVigna and Malmendier, 2004), other analyses consider "nudges" that are intended to improve the agent's decision making and welfare (e.g., Thaler and Sunstein, 2008).

A key topic in the emerging literature on behavioural contracting is loss aversion, wherein agents evaluate outcomes relative to a reference point (for reviews, see Rabin, 1998; DellaVigna, 2009; and Köszegi, 2015). Loss aversion models postulate that gains increase utility less than comparable losses decrease utility. The notion of loss aversion is formalized in prospect theory (Kahneman and Tversky, 1979), and is associated with well-known behavioural

anomalies such as the endowment effect (Thaler, 1980), status quo bias (Samuelson and Zeckhauser, 1988), and diverging values of willingness to pay and accept (Kahneman et al., 1990; Hanemann, 1991).[1]

With theory in hand, one might wonder if loss aversion can be leveraged to increase (labor) productivity? In an interesting field experiment, Abeler et al. (2011) provide suggestive evidence in the affirmative. They exogenously shift agents' expectations about the remuneration they may receive, and find that effort adjusts accordingly – in line with reference-dependent utility theory. Indeed, reference points can also be shifted using the *timing* of the earnings – before or after the (experimental) task. Bonuses are typically paid after a specific productivity target has been met. Alternatively, one may offer money up front, which is subsequently taken away if the agent fails to meet the target. The efficacy of the latter incentive scheme (typically referred to as a claw-back or penalty regime) was first tested in a field experiment due to Hossain and List (2012). Collaborating with a Chinese electronics company, they implement a simple framing experiment where a random subsample of workers is promised a bonus on top of their salary, to be paid at the end of the week, if a productivity threshold is met. Others were given a salary enhancement, which was claimed back at the end of the week in case of low productivity. Hossain and List (2012) find that even such a weak framing treatment raises productivity of teams of workers relative to the economically isomorphic bonus treatment, framed in a conventional sense.

Fryer et al. (2018) complement this work by providing financial incentives to school teachers to increase productivity as measured by the performance of their students. While conventional bonuses fail to increase teacher performance considerably, leveraging loss

[1] More recently, reference-dependent utility has been introduced in models explaining issues such as wage setting and bonus contracts (de Meza and Webb, 2007; Herweg et al., 2010), selling behavior on the housing market (Genesove and Mayer, 2001), product demand (Herweg and Mierendorff, 2013), and labor supply (Camerer et al., 1997; Goette et al., 2004; Fehr and Goette, 2007, Crawford and Meng, 2011; Cohn et al., 2017).

aversion via a penalty regime is found to be effective in the short run. Yet there were no significant effects in a follow-up wave of the experiment, suggesting the loss aversion effect may taper off over time. Levitt et al. (2016) incentivize students, rather than teachers, and find that the claw-back incentive regime outperforms a conventional bonus regime, awarding good performance on tests ex post.[2]

While the literature focuses on how loss aversion promotes effort for exogenously-imposed incentive regimes, an emerging literature considers the demand for incentive regimes – allowing for a more complete consideration of potential welfare effects. One major question is whether agents voluntarily choose dominated contracts – contracts penalizing underperformance but providing no extra rewards for sufficiently high performance. In conventional agency models, such contracts would not be selected because they imply additional risk for the agent for which she demands compensation. However, the conventional model ignores self-control problems. Workers who are aware of their self-control issues (or who are "sophisticated"), may rationally prefer to be exposed to sharp incentives to "tie themselves to the mast." In other words, contracts based on penalizing underperformance may have value as a commitment device.[3] Recent work by Kaur et al. (2015) and Beshears et al. (2015) suggests many agents are willing to accept externally enforced restrictions to incentivize their own (future) behaviour.

The commitment value of dominated contracts has implications for the case of bonus versus penalty regimes. Standard theory predicts that loss averse agents will prefer bonus contracts over penalty contracts when the underlying pay-offs are identical. Earlier work by,

[2] Loss aversion does not always appear to affect behavior. Hossain and List (2012) fail to document that the claw-back regime enhances the productivity of individual (as opposed to teams of) workers, and List and Samek (2015) do not find that loss aversion helps dieticians to affect children's food choice. Understanding the boundary conditions of loss aversion remains an important yet under-researched line of work.

[3] Following Bryan et al. (2010, p.672) we define a commitment device as "an arrangement entered into by an individual with the aim of helping fulfill a plan for future behaviour that would otherwise be difficult owing to intrapersonal conflict stemming from, for example, a lack of self-control."

for example, Luft (1994), Hannan et al. (2005), and Brink and Rankin (2013) indeed established that subjects prefer bonus contracts over claw-back contracts in experimental settings. But, if penalty regimes enable agents to *commit* to reaching performance targets, they may be preferred by sophisticated agents nonetheless. Imas et al. (2016) conjecture that this mechanism explains why agents prefer penalty regimes over bonus regimes in their lab experiment (but see De Quidt 2018 for diverging evidence).

This paper seeks to contribute to the literature in three ways. First, we implement a field experiment to probe the robustness of earlier findings and test whether a claw-back regime induces higher effort levels than an otherwise identical bonus regime. In spite of accumulating evidence, this issue remains unclear and controversial.[4] Our field experiment mimics the workplace and is set in a novel context (peri-urban Uganda) using a large sample of "non-standard" subjects (Henrich et al. 2010). Second, as in Imas et al. (2016) and De Quidt (2018), we allow subjects to choose between the two regimes in a two-stage design, and ask when subjects self-select into regimes incentivized via the claw-back. A natural question is whether previous experience with the claw-back affects one's preference for the mechanism. Third, we examine the welfare implications of introducing a claw-back incentive regime.

We present several insights. First, we document large productivity increases caused by the claw-back. On average, across the two tasks in our field experiment, productivity increases by 20%. Second, after being exposed to the claw-back incentive scheme, a substantial share of our subjects act in accord with a model of a sophisticated agent, who has "learned" how to leverage the claw-back to commit to supplying higher effort levels in a subsequent task. Such commitment is not optimal for all workers, however, depending on the combination of

[4] For example, DellaVigna and Pope (2017) asked a sample of economic experts to forecast how productivity in a bonus frame would compare to productivity in a penalty frame, and find that forecasted productivity levels for an equal sized bonus/penalty are not statistically different.

behavioural biases affecting the worker's job performance, her level of sophistication, and the nature of the task at hand. We develop a dual-self principal-agent model to derive under what circumstances subjects prefer the claw-back, focusing on the role of the claw-back as a self-commitment device. The theory predicts, consistent with our data, that there should be little demand for the claw-back, even among sophisticated workers, when the task is not tedious. Third, we demonstrate considerable heterogeneity in valuation for the claw-back. Some workers self-select out of the claw-back regime even at large expected financial cost, whereas others self-select into the claw-back for minimal expected gains.

A final important result is that even accounting for those workers who select out of the claw-back regime, we observe very modest negative effects of the claw-back. Indeed, the utility losses from the claw-back are approximately offset by the value of commitment for the average worker.[5] Accordingly, it does not appear that any productivity gains experienced by the firm are obtained at the expense of worker utility loss. This result provides an initial indication of the potential efficacy of using behavioural insights, such as loss aversion, to encourage workers to put forth higher effort levels. We also argue, based on the theoretical model, that the claw-back can be used by employers as a screening device to identify different types of workers in the population.

The remainder of our study is organized as follows. In Section 2, we illustrate the workings of the claw-back regime by means of a simple model, define concepts such as sophistication and commitment relevant to our context, and present the hypotheses the model gives rise to. In Section 3, we introduce our field experiment, describe the data, and outline our

[5] Our analysis focuses on self-selection into tasks, following DellaVigna et al. (2012), who also link theory to an experiment to allow measurement of the intervention's impact on welfare. DellaVigna et al. use randomization to identify behavioral parameters in their models, but unlike our work, do not measure the welfare effects of a nudge for the worker. Alternative approaches to measuring welfare counterfactuals include the Bernheim-Rangel criterion (Chetty et al., 2009), prodding inert people to make a choice (Carroll et al., 2009), or asking for the WTP for a nudge (Allcott and Kessler, 2019).

identification strategy. In Section 4, we present the results on productivity, selection, and welfare. As a simple robustness analysis we also ask whether cognitive skills (proxied by education levels) affect the ability of participants to recognize the potential commitment value of the claw-back. Section 5 concludes.

2. Theory and hypotheses

Köszegi (2015) identifies four main insights from behavioural economics that have prominently found their way into economic analysis: loss aversion, present bias (e.g. hyperbolic discounting), inequity aversion, and overconfidence. The most common approach is to augment standard economic models with psychological foundations, and include the above-mentioned preferences or beliefs into formal models of decision making and contracting. Dual-self (principal-agent) models are developed in which oftentimes the "implementing-self" (the agent) behaves in accordance with a psychologically-based model, and in which the "planning-self" (the principal) is a rational utility (or profit) maximizer. Models based on psychological foundations may produce outcomes that diverge from standard (micro-economic) predictions, affecting both overall efficiency and the distribution of the surplus.

In Section 2.1 we develop a model of evolving sophistication and derive consequences for productivity and self-selection into incentive regimes. This model guides the development of our hypotheses, presented in Section 2.2, to be tested in the experiment.

2.1 The model

We formulate and solve a model in the spirit of Thaler and Shefrin (1981), O'Donoghue and Rabin (1999) and Fudenberg and Levine (2006). We assume that workers may suffer from both loss aversion and a self-control problem. Lack of self-control means that workers often do not work as hard as they themselves prefer (for reviews, see Frederick et al. (2002), and DellaVigna (2009)), especially when the task at hand is tedious. *Ex ante*, a worker prefers to

exert high effort and work hard. But comes the time to work he is tempted to procrastinate or shirk. Bryan et al. (2010) refer to this as outcomes when beliefs regarding costs and benefits of specific activities are time-varying.[6] We capture the distinction between the "planner" and "the implementer" within the worker as a "dual self", and develop a simple intra-person principal-agent model.[7] This is consistent with McIntosh (1969), as cited by Fudenberg and Levine (2007), who wrote that "the idea of self-control is paradoxical unless it is assumed that the psyche contains more than one energy system, and that these energy systems have some degree of independence from each other."[8]

Assume each worker *i* consists of two personalities: the rational and reflexive (but possibly naïve) planner-self, called the "principal", who aims to maximize overall utility, as well as an implementer-self (the "agent") that suffers from behavioural biases. The decision problem we study may be regarded as a game between the biased "implementer self," responsible for deciding about effort levels and the rational "planner-self," responsible for selecting the incentive regime in which the former works. The underlying idea is that the rational principal might try to manipulate her agent's behaviour by choosing a certain incentive structure. Our line of reasoning deviates from the pioneering work of O'Donoghue and Rabin (1999) in various respects. They assume a game between a principal and agent, where the agent "suffers" from one behavioural bias – a self-control problem. The agent overestimates the costs associated with supplying effort, relative to the future benefits of successfully completing the task. The principal is either "sophisticated" or "naïve" depending on whether she knows the utility function of her agent. We, however, assume the agent potentially suffers from *two*

[6] In addition to the dual-self model we develop, commitment failures may also be explained by quasi-hyperbolic discounting (Laibson (1997) or by choice-set-dependent utility (Gul and Pesendorfer, 2001, 2004).
[7] See Thaler and Shefrin (1981) for a pioneering contribution focusing on the consumption-savings problem, and Fudenberg and Levine (2006) for a more general treatment.
[8] Fudenberg and Levine (2006) also cite recent evidence from MRI studies suggesting that different parts of the brain are involved in long-term planned behavior and in short-term impulsive behavior. For a discussion of additional neuroscientific evidence, refer to Bryan et al. (2010) who cite pioneering work of McClure et al. (2004) and Shiv and Fedorikhin (1999).

behavioural biases (self-control and loss aversion). Moreover, while all principals know their agents' propensity to shirk (lack of self-control), we assume not all principals know whether their agent is loss averse. In our setting—adult workers—it seems realistic to assume that workers have had ample opportunity to learn about their shirking or procrastination behaviour. In contrast, most planner-selves are not familiar with claw-back incentives (described below), and it seems reasonable to assume that they may be unaware of their implementing-selves' degree of loss aversion. Naïve principals know that their agents suffer from a lack of self-control but mistakenly believe that they are not loss averse. Sophisticated principals, through experience, know both their agents' self-control issues and their level of loss aversion. We vary experience with the claw-back in the experiment, and test whether experience affects the principal's level of sophistication and choice of payment regime.

2.1.1 The planning-self's maximization problem

Assume worker *i* is offered to perform a specific task, task *j*. The planner-self of worker *i* maximizes a conventional pay-off function based on material outcomes and effort cost:

$$V_i = p(e_{ij})R - c(e_{ij}) \quad (1)$$

where $p(e_{ij})$ measures the probability that the worker meets a performance threshold so that she earns payment *R*. Effort allocated to the task is denoted by e_{ij}, and we assume $p'(e_{ij}) > 0$ and $p''(e_{ij}) \leq 0$. Effort costs are denoted by $c(e_{ij})$, with $c'(e_{ij}) > 0$ and $c''(e_{ij}) > 0$. From the perspective of the planning-self, the optimal effort level e_{ij}^* solves $c'(e_{ij}) = p'(e_{ij})R$. While the planning-self cannot choose effort herself (this is decided by the agent, or implementing-self), she chooses the incentive regime within which her agent works. Suppose there are two incentive regimes:

- A *bonus regime*, where the worker receives *R* if productivity exceeds a threshold level *T* determined by median productivity in a reference population; and
- A *claw-back regime*, where the worker receives *R* before commencing the task, and has to return this payment in case productivity falls short of the same threshold *T*.

Across regimes, the financial rewards are the same and only depend on whether the worker meets a performance threshold. The regimes differ in the timing of provision of the reward. For workers without reference-dependent preferences this distinction is immaterial. Loss-averse agents will, however, experience a loss in utility when they have to return the payment, and this affects effort.

2.1.2 The implementing-self's effort decision

Assume the implementing self of worker *i* has the following utility function, which captures both self-control bias and loss aversion (or reference-dependent utility):

$$u_i = \begin{cases} w_i - \alpha_{ij} c(e_{ij}), & if \;\; w_i \geq r_i \\ w_i + \theta_i (w_i - r_i) - \alpha_{ij} c(e_{ij}), & if \;\; w_i < r_i \end{cases} \tag{2}$$

where $w_j \geq 0$ are the actual earnings worker *i* receives when successfully completing task *j*, $r_i \geq 0$ is the reference value used by the agent to which he compares his earnings, e_{ij} is effort, and (θ_i, α_{ij}) are parameters. Parameter α_{ij} captures a potential lack of self-control of the implementing-self of worker *i* while performing task *j*. We follow Bryan et al. (2010) and argue that lack of self-control is also task-specific, α_{ij}, with *j* indexing tasks. For some tasks it is easy to maintain a steady effort level, while for other tasks this is difficult because they are tedious or uninteresting. In what follows we assume $\alpha_{i1} > \alpha_{i2}$ if task 1 is more "tedious" than task 2. Similar to O'Donoghue and Rabin (1999), the agent's lack of self-control is captured by $\alpha_{ij} \geq$

1, which affects the agent's relative valuation of earnings and effort costs, as well as the optimal amount of effort as perceived by the agent.

Agent i is loss averse if $\theta_i \geq 0$. Loss aversion implies earnings (outcomes) below a reference value induce a loss in utility, and we assume the magnitude of the loss is linear in the distance between the realized outcome and the reference value. Specification (2) implies the disutility caused by an outcome *w* falling short of reference value *r* is (weakly) larger than the utility gain caused by an outcome exceeding that reference value by the same quantity.

Worker *i* is invited to engage in task *j*, which has two possible outcomes. If production equals or exceeds an exogenous (and unknown) threshold *T*, then she receives a payment equal to $R > 0$. When production is below the threshold, she receives nothing, so that $w_i = \{0, R\}$. Production in task *j* is a positive function of effort, e_{ij}, and the probability of meeting threshold *T* is denoted by $p(e_{ij})$, as discussed above.

We consider two incentive regimes with different reference values. In a *bonus regime*, the agent's reference value equals zero ($r_i = 0$) and the agent maximizes the following function:

$$max_{e_{ij}}\, u_i = p(e_{ij})R - \alpha_{ij}c(e_{ij}). \qquad (3)$$

The agent's optimal effort level, e_{ij}^B, implicitly solves $c'(e_{ij}) = p'(e_{ij})R/\alpha_{ij}$. While e_{ij}^B is not a function of θ_i, it is a function of α_{ij}. Because $p" \leq 0$ and $c" > 0$, the more the agent "inflates" the cost associated with providing effort for a tedious task, the lower the level of effort chosen – the self-control problem ($de_{ij}^B/d\alpha_{ij} < 0$).

In a *claw-back regime*, instead, the agent's reference value (might) equal the ex-ante payment ($r_i = R$). Prior experiments (see List, 2003, 2004) have revealed that ex-ante transfers create sentiments of ownership, so the agent maximizes:

$$max_{e_{ij}}\, u_i = p(e_{ij})R - \left(1 - p(e_{ij})\right) R\theta_i - \alpha_{ij} c(e_{ij}). \quad (4)$$

In a claw-back regime the agent chooses effort level e_{ij}^C, which solves $c'(e_{ij}) = p'(e_{ij})(1 + \theta_i)R/\alpha_{ij}$. It is clear that e_{ij}^C is a function of both α_{ij} and θ_i. In addition, for $\alpha_{ij} \geq 1$ we have $e_{ij}^C(\alpha_{ij}, \theta_i) > e_{ij}^B(\alpha_{ij})$ if $\theta_i > 0$: loss averse agents "work harder" to avoid the loss associated with giving up the payment, and the difference in effort levels is larger the higher is θ_i.[9]

2.1.3 The planner-self's welfare levels under the two regimes

Given the agent's effort problem discussed above, what incentive regime should the worker's planning-self choose? To answer this question we must evaluate the consequences of the agent's effort decisions for the worker's welfare, *as perceived by the principal*. Welfare as perceived by the principal is highest if the agent's effort supply is as close as possible to e_{ij}^*.[10]

So how does the principal's welfare vary with the payment regime within which her agent implements the task, given the agent's lack of self-control and loss aversion? From the principal's perspective the agent will not supply enough effort in a bonus regime if he lacks self-control $(e_{ij}^B < e_{ij}^*)$. In a claw-back regime the agent can work "too hard" $(e_{ij}^C > e_{ij}^*)$ or not hard enough $(e_{ij}^C < e_{ij}^*)$, depending on parameters and on the type of task. Specifically, for $(1 + \theta_i) > \alpha_{ij}$ the agent will supply too much effort $(e_{ij}^C > e_{ij}^*)$. In contrast, for $(1 + \theta_i) < \alpha_{ij}$ the agent supplies less effort than is optimal, but still more than under the bonus scheme $(e_{ij}^B < e_{ij}^C < e_{ij}^*)$. This can be seen as follows.

[9] In studies of the claw-back that involve a significant delay in repaying the money (e.g. as in Fryer et al., 2018) an additional effect may be relevant. Early payment of the bonus may relax a binding financial constraint that could enable the subject to perform better (via enhanced nutrition, say, or complementary inputs privately purchased; see for example Mani et al. (2013). In our experiment, the time difference between receiving the bonus in the bonus and claw-back regimes is maximally just 30 minutes so that money cannot be spent.

[10] It is common to assume that the principal maximizes her own pay-off function, and does not attach any weight to her agent's welfare. We will adopt this convention, but also observe that the principal and agent are the same real person of 'flesh and blood.' It is therefore not obvious that completely disregarding the agent's utility is necessarily optimal – it may make sense for the principal to avoid outcomes that would make her deeply unhappy in her capacity as the agent. We will return to this issue below when we present our experimental hypotheses.

Let us use $V^* = V(e_i^*)$ to denote the maximum welfare the principal can attain (see (1), and $V^B = V\left(e_{ij}^B(\alpha_{ij})\right)$ and $V^C = V\left(e_{ij}^C(\alpha_{ij}, \theta_i)\right)$ to denote the welfare levels obtained by the principal if she selects the agent into the bonus and the claw-back regime, respectively. We now pose the following lemma:

Lemma 1: *For any $\theta_i > 0$, there is a critical level of α_{ij}, $\overline{\alpha}_j(\theta_i)$, where the principal's welfare is equally high if the agent works under either the bonus regime or the claw-back regime. The principal's welfare is higher in the bonus regime than in the claw-back regime for $\alpha_{ij} < \overline{\alpha}_j(\theta_i)$, and the opposite holds if $\alpha_{ij} > \overline{\alpha}_j(\theta_i)$.*

Proof: Suppressing subscripts to avoid clutter, from the above first-order conditions it immediately follows that $V^B(\alpha) = V^*$ if $\alpha = 1$, and $V^C(\alpha, \theta) = V^*$ if $\alpha = 1 + \theta$. Next, $\frac{dV^B}{d\alpha} = [p'R - c'] \frac{de^B}{d\alpha} = \frac{(p'R - c')c'}{p"R - \alpha c"} < 0$ because $e^B(\alpha) < e^*$ for all $\alpha > 1$. Similarly, we have $\frac{dV^C}{d\alpha} = [p'R - c'] \frac{de^C}{d\alpha} = \frac{(p'R - c')c'}{(1+\theta)p"R - \alpha c"}$. Because $e^C(\alpha, \theta) > e^*$ if $\alpha < 1 + \theta$ and $e^C(\alpha, \theta) < e^*$ if $\alpha > 1 + \theta$, we have $\frac{dV^C}{d\alpha} > (<) 0$ if $\alpha < (>) 1 + \theta$.

Combining (i) $V^B(1, \theta) = V^C(1 + \theta, \theta) = V^*$, (ii) $\frac{dV^B}{d\alpha} < 0$ for all $\alpha > 1$ and (iii) $\frac{dV^C}{d\alpha} > 0$ for $1 \leq \alpha < 1 + \theta$, there exists a critical level of α, $1 \leq \overline{\alpha}(\theta) < 1 + \theta$, such that $V^C(\alpha, \theta) < V^B(\alpha)$ for $\alpha < \overline{\alpha}(\theta)$ and $V^C(\alpha, \theta) > V^B(\alpha)$ for $\alpha > \overline{\alpha}(\theta)$. ∎

Lemma 1 is illustrated in Figure 1, which plots the principal's welfare for a range of (lack of) self-control values of the agent for the task at hand (α_{ij}), where the agent's loss aversion may be high or low (θ^H, θ^L, with $\theta^H > \theta^L$), if the agent works under a bonus regime or under a claw-back regime. Loss aversion does not affect effort when it is performed under a bonus regime. Here, the agent supplies a level of effort that maximizes the principal's welfare if $\alpha_{ij} = 1$ (and hence $V^B(\alpha_{ij}) = V^*$ if $\alpha_{ij} = 1$). He will put in less effort when $\alpha_{ij} > 1$. The

difference between e_{ij}^{*} and $e_{ij}^{B}(\alpha_{ij})$ increases with α_{ij}, and the principal's welfare level $V^{B}(\alpha_{ij})$ is a monotonically decreasing function of α_{ij}.

<Insert Figure 1 about here>

Effort of the agent does depend, however, on the degree to which he suffers from loss aversion if he works under a claw-back regime. Hence, the principal's welfare levels differ too. If a loss-averse agent who does not suffer from a lack of a self-control for a specific task (i.e., $\alpha_{ij} = 1$) works under the claw-back regime, his effort level is too high from the principal's perspective. The difference will be larger if the agent is more loss averse ($V^{C}(1, \theta^{H}) < V^{C}(1, \theta^{L}) < V^{*}$). For $\alpha_{ij} > 1$, the difference between e_{ij}^{*} and $e_{ij}^{C}(\alpha_{ij}, \theta_i^k)$, $k = \{L, H\}$, first decreases, becomes zero (at $\alpha_{ij} = 1 + \theta_i^k$), and then becomes more and more negative. This means that $V^{C}(\alpha_{ij}, \theta_i^k)$ is a hump-shaped function of α_{ij}, resulting in $V^{C}(\alpha_{ij}, \theta_i^k) = V^{*}$ at $\alpha_{ij} = 1 + \theta_i^k$. Combining, for given θ_i the principal prefers to select his agent into the claw-back regime (as opposed to the bonus regime) if $\alpha_{ij} > \bar{\alpha}_j(\theta_i)$. As shown in Figure 1 we have $d\bar{\alpha}_j(\theta_i)/d\theta_i > 0$, and the optimal choice for the principal depends on the combination of α_{ij} and θ_i.[11]

So what regime will the principal select for their agent to work in? All planner-selves are aware of the extent to which their agents suffer from a lack of self-control; life offers plenty of situations to provide this insight. Most planner-selves, however, are not familiar with claw-back incentives, and it seems reasonable to assume that they may be unaware of their implementing-selves' degree of loss aversion. The outcome depends on whether the principal is sophisticated, or naïve. We assume the following.

[11] Figure 1 is derived using $p(e) = \frac{e - e^{Min}}{e^{Max} - e^{Min}}$ and $c(e) = 0.5e^2$, with $e^{Max} = 2, e^{Min} = 0$, and using $R = 2\sqrt{2}, \theta^{L} = 0.2, \theta^{H} = 0.4$. Note that R is chosen such that $V^{*} = 1$. Solving, we have $\bar{\alpha}(\theta) = (2 + \theta)/2$. Hence $\partial\bar{\alpha}(\theta)/\partial\theta > 0$, $\bar{\alpha}(\theta^{L}) = 1.10$ and $\bar{\alpha}(\theta^{H}) = 1.20$. Alternatively, we can write $\bar{\theta}(\alpha) = 2(\alpha - 1)$, and then the principal is better off under the claw-back than in the bonus regime if $\theta < \theta(\alpha)$. If $\theta > \theta(\alpha)$, the agent works too hard under the claw-back, and the principal's welfare is higher in the bonus regime.

1. *Sophisticated principals* know the extent to which their implementing-selves suffer from loss aversion (captured by θ_i). They will choose the claw-back if e_{ij}^{C} yields greater net pay-offs for the principal (as evaluated by (1)) than e_{ij}^{B} (see below);
2. *Naïve principals* mistakenly believe their agent is not loss averse, or that $\theta_i = 0$, and are always indifferent between the bonus and claw-back regime.

The bonus scheme is more standard than the claw-back scheme, and hence we assume that naïve principals weakly prefer the bonus scheme due to their past experiences. Sophisticated planner-selves, instead, anticipate that the claw-back regime disciplines their loss-averse implementer-selves, inducing them to work harder and increasing the odds of actually earning the reward. They thus recognize that loss aversion can be leveraged to overcome the self-control problem. The sophisticated principal will select her agent in the bonus regime for $\alpha_{ij} < \bar{\alpha}_j(\theta_i)$; if not, the sophisticated principal prefers to select the agent into the claw-back regime. Hence:

Corollary 1: *Naïve principals are indifferent between the bonus and the claw-back incentive regime. Sophisticated principals prefer the claw-back to the bonus if the ratio α_{ij}/θ_i is sufficiently high.*

Proof: This follows immediately from Lemma 1. ∎

The ratio α_{ij}/θ_i depends on the type of task *j*; the more tedious the task, the more severe the agent's lack of self-control. For non-tedious tasks, maintaining a sufficiently high level of effort is less difficult, and sophisticated principals should "switch" to the bonus contract. This gives rise to Corollary 2:

Corollary 2: *For any $\theta_i > 0$, the principal's welfare is more likely to be higher in the bonus scheme than in the claw-back scheme if the task is non-tedious.*

Proof. By assumption, non-tedious tasks are tasks for which α_{ij} is close to 1 and shirking does not occur. That means that $P\left(\alpha_{ij} < \bar{\alpha}_j(\theta_i)\right)$ is larger the smaller is α_{ij}, and hence the more likely it is that a sophisticated principal will select her agent into the bonus regime. ∎

Note the following implication. For given distribution of θ_i among a pool of workers, the ratio α_{ij}/θ_i is low if the task is non-tedious, and hence (i) the share of workers selecting into the claw-back regime should be lower too (because only the not-so-loss averse or the ones facing the strongest levels of self-control issues select into the claw-back); and (ii) the average productivity differences between those selecting into the claw-back and into the bonus are smaller because those with the lowest propensity to shirk prefer to select into the bonus regime, to avoid ending up working too hard in the claw-back regime.

2.1.4 Experience and sophistication

Where does sophistication come from? We follow the literature and assume that principals may receive a signal about the nature of the agent's loss aversion after observing his behaviour. Principals learn nothing new in a bonus regime, but may glean valuable information about their agent's preferences by observing him supply effort in a claw-back regime. Specifically we assume the following about experience and sophistication:

1. Observing the agent under a bonus contract does not allow a naïve principal to learn that $\theta_i > 0$.
2. However, observing the agent under a claw-back regime is informative for naïve principals – the agent works harder than expected, revealing $\theta_i > 0$ so that the principal becomes "sophisticated."

2.1.5 The welfare effects of the claw-back mechanism

How can we infer the welfare effects of the claw-back for different types of workers? When offered the option to work under a claw-back regime or for a fixed wage, the planner-self selects the preferred option – comparing her expected welfare under both regimes.

When comparing the opportunity to perform a task under a claw-back regime or under a fixed wage (W), sophisticated workers are better off under the claw-back if $V_S^C(\alpha_{ij}, \theta_i) \geq W$. This yields the following participation constraint:

$$p(e_{ij}^C(\alpha_{ij}, \theta_i))R - c\left(e_{ij}^C(\alpha_{ij}, \theta_i)\right) \geq W. \quad (5)$$

In terms of Figure 1, the fixed wage W can be introduced as a horizontal line (with vertical intercept W).[12] We know that for any θ_i, $V_S^C(\alpha_{ij}, \theta_i)$ is an inverted U-shaped curve that reaches its maximum where $\alpha_{ij} = 1 + \theta_i$. The horizontal line may intersect the inverted U-shaped $V_S^C(\alpha_{ij}, \theta_i)$ curve once, twice, or zero times, depending on the relative value of W:

- *If W is (very) large*, above the top of the inverted U-shaped $V_S^C(\alpha_{ij}, \theta_i)$ curve, then the horizontal line does not cut the claw-back pay-off curve and all workers are better off under the fixed wage.
- *If W is of intermediate size*, located between the vertical intercept of the $V_S^C(\alpha_{ij}, \theta_i)$ curve and its peak, then the horizontal line cuts the claw-back pay-off curve twice and there are three types of workers: (i) low-shirkers who are better off under a fixed wage, (ii) intermediate-shirkers, who are better off under the claw-back, and (iii) high-shirkers, who are again better off under a fixed wage.

[12] We assume that with a fixed wage the agent rationally chooses to not supply any effort, such that $c(e_{ij})=0$.

- *If W is low*, located below the vertical intercept of the $V_S^C(\alpha_{ij}, \theta_i)$ curve, then the horizontal line cuts the claw-back pay-off curve once. Low-shirkers and intermediate-shirkers are better-off under the claw-back, and extreme shirkers prefer the fixed wage.

The welfare effect of introducing the claw-back therefore varies across individuals, depending on their behavioural preferences (θ_i,α_{ij}) and the opportunity cost (or the level of fixed wage *W*).

2.2 Hypotheses

Having developed the model, we now present our hypotheses. When workers are randomized into regimes, self-control and loss aversion are orthogonal to treatment status, and distributions of these behavioural traits should be identical across regimes. We expect that a non-negligible share of our subject pool is loss averse, and work extra hard to avoid the penalty in a claw-back regime. Hence, we state the following hypothesis.

Hypothesis 1: *Average productivity for a population of workers is higher in a claw-back regime than in a bonus regime.*

Regarding the workers' choice of working regime, we distinguish between naïve and sophisticated planner-selves. As stated, naïve planner-selves mistakenly believe their implementer-self is not loss averse and therefore indifferent between the claw-back and bonus regime. Assuming that in this case the planner-selves choose the regime that is more familiar, naïve planner-selves will select their agents into the bonus regime. In contrast, a sophisticated planner-self recognizes that loss aversion can be leveraged to reduce shirking, and may prefer her implementing-self to work in a claw-back regime. We exploit experimentally-induced differences in experience with the claw-back as a proxy for sophistication, and specify the following hypothesis.

Hypothesis 2*: For given distributions of self-control and loss aversion in the population, experience with the claw-back fosters sophistication and thereby increases demand for the claw-back regime.*

Sophisticated principals will only choose the claw-back if a commitment device has value – when the task at hand is sufficiently tedious. For short or non-tedious tasks, where most workers do not shirk, additional motivation for the implementing-self to supply effort does not make the planning-self better off. The reverse may be true, if too much effort is supplied. For tedious tasks, instead, implementer-selves lacking self-control are prone to shirk, so leveraging loss aversion may motivate the implementer-self to choose an effort level that is closer to the planning-self's optimum. We exploit experimentally-induced differences in "tediousness of the task," and specify the following hypothesis.

Hypothesis 3*: For given distributions of loss aversion and sophistication in the population, the share of workers choosing the claw-back regime is larger the more likely it is that workers suffer from self-control problems because they have to perform a "tedious task".*

Next, we take a first step towards considering the welfare effects of the two incentive regimes. Based on the reasoning above, the net welfare effect for a group of workers will depend on the commitment value of the claw-back, relative to the cost of over-supplying effort, and on the population share of experienced principals in the population.

Hypothesis 4*: The net welfare effect of exogenously introducing a claw-back incentive regime is ambiguous.*

3. Experimental design and data

To test our predictions we designed and implemented a field experiment in the suburbs of Kampala, Uganda. In total, 1200 subjects participated in our field experiment: 200 in the pilot

phase and 1000 in our four key treatments, which we label A,B,C,D. Each experimental treatment consisted of two parts, and in each part the subject was invited to participate in a real-effort task – 30 minutes of producing envelopes (Task 1) followed by sorting beans (Task 2). We use *i* to denote subjects, *j* to denote tasks, and *z* to denote treatments. We selected folding envelopes and sorting beans because output Q_{ijz} (i=1,...,1000; j=1,2; z=A,B,C,D) was easy to measure for these tasks, and because we can control for quality of the output produced. We only accepted envelopes that were strong enough to withstand firm shaking when filled with coins, and only accepted bags of beans that were perfectly sorted (on the basis of color).

Before starting the field experiment, we implemented a pilot study involving 200 subjects to learn about the distribution of productivity. This enabled us to set realistic thresholds for the treatment arms in the main experiment. In the pilot, we asked subjects to fold and glue envelopes for 30 minutes, and to sort beans for an equal amount of time. Using a between-subject pilot design, we offered both piece rate compensation as well as a fixed wage in the pilot, and measured output. Based on performance in the pilot, we set the threshold for payment in the treatments implementing the bonus and claw-back regimes in the main experiment. Specifically, we set the threshold for treatments A, B and D at the median output levels in the pilot treatments: T_1 = 18 folded envelopes and T_2 = 340 grams of sorted beans. The actual thresholds were not disclosed until after the subjects had completed the tasks, and the only information provided *ex ante* was that the thresholds were set equal to the median level of productivity in pilot sessions.[13]

[13] Subjects were not informed about the exact level of the performance threshold in any of the treatments. Not informing subjects about the threshold results in a range of beliefs about the required amount of effort to receive the reward (or not to lose it), but we have no reason to assume these expectations will systematically vary across treatments. The main reason why we decided not to disclose the threshold is for statistical reasons. If the threshold is known, the only outcome measure we have is whether a subject managed or failed to reach the threshold – few subjects would supply positive (costly) effort after reaching the threshold, and others might stop trying if they felt the threshold was out of reach. We believe not disclosing the threshold enables us to better measure production across treatments (see also Imas et al., 2016).

<< Insert Table 1 about here >>

Next, turn to our two-task experiment, summarized in Table 1. The protocol, which was translated into Luganda, is provided in the only replication package. Our 1000 individuals received a show-up fee of UGS 2000.[14] Since we wish to vary experimentally the level of experience (as a proxy for sophistication) of the principals, we *exogenously* allocate subjects to perform their first task in either a bonus regime or a claw-back regime. A random sub-sample of 200 subjects was allocated to Treatment A – performing Task 1 under a bonus regime. These workers were promised a payment of UGS 2500 if they would meet the (undisclosed) productivity threshold -- completing 18 envelopes, or more. Subjects received instructions about the task, were shown the money in small envelopes long enough to enable them to verify the content, and were informed about the conditions under which they would receive the payment of UGX 2500.

The remaining 800 subjects were randomized into Treatments B, C, and D, and worked under the claw-back regime for Task 1. These workers received the same information about the task at hand, and received their payment of UGX 2500 "up front". They were invited to count the money and to place the envelope with cash on the table in front of them with the lid open (so the money was always in view). If they failed to meet the undisclosed productivity threshold of completing 18 envelopes within the thirty minute period, then they had to return the envelope with cash to one of the experimenters. Comparing output across the treatment arms allows us to assess whether the claw-back regime yields higher productivity – testing hypothesis 1.

We assume that workers performing Task 1 under a bonus regime (Treatment A) learned little about the extent to which they suffer from loss aversion. In contrast, workers in the other three treatments (B,C,D) gained experience with the claw-back during the first task, so that each

[14] 1 USD = 2800 UGS, and the average daily wage was about 6000 UGS or USD 2.14.

worker's planner-self received a signal about their implementing-self's level of loss aversion. The model in Section 2.1 is based on the assumption that this signal fosters sophistication.[15]

After Task 1 participants were given a short break, during which they learned whether they had reached the threshold. Those who reached the threshold in Treatment A received their payment envelope with UGS 2500, and those who failed to reach the threshold in Treatments B,C,D were forced to give theirs back. All subjects complied, albeit sometimes grudgingly. Settling the payments for Task 1 before the start of Task 2 should help mitigate concerns that having the payment envelope on one's desk rather than at the experimenter's desk resulted in differential levels of trust in actual payment.

After the break, subjects were informed that they would perform a second task (bean sorting), and that they would have the opportunity to choose the payment regime under which they wished to perform that task. Subjects were reminded of the details of the payment regime that was in place for their Task 1, and were explained the details of the other regime. Subjects in Treatment A received additional information about the claw-back regime, and subjects in Treatments B-D were informed about the details of the bonus regime.

Upon completion of the instructions, subjects were invited to *choose* the incentive regime under which they wanted to implement Task 2. Task 2 in Treatments A and B lasted 30 minutes, which presumably makes bean sorting quite tedious. We conjecture that it is more difficult to maintain self-control over effort for tedious tasks that last longer.[16] Comparing self-

[15] Whether exposure to a commitment device in a thirty-minute task is enough to learn about the value of commitment, was an open question at the time we designed the experiment. Kahneman et al. (1990) and List (2003, 2004) show that the valuation of coffee mugs and sportscard memorabilia vary instantaneously with the change in ownership, suggesting that the endowment effect arises instantaneously. Closer to our paper, Augenblick et al. (2015) confront subjects with their own propensity to procrastinate, and subsequently offer them a commitment device. They find that people learn quickly, and start demanding (costly) commitment.

[16] In the language of our model in Section 2.1, bean sorting for half an hour is a "high-α" task.

selection of subjects in Treatments A and B, we can test hypothesis 2 and assess whether experienced participants are more likely to choose the claw-back regime.

While our preferred channel linking Task 1 and the choice for incentive regimes in Task 2 is a difference in sophistication (due to experience), we acknowledge that it is difficult to rule out other channels. As argued above, differences in trust may affect this choice as well even if our experiment was designed to minimize such concerns (the payment was always in view of the subjects – both in the bonus and claw-back treatment – and we paid the subjects for their performance immediately after Task 1).

To further probe this issue, and test hypothesis 3, we experimentally vary the importance of demand for self-control. In Treatment C, experienced subjects are offered a choice between the claw-back and bonus regime, but here the second task lasts only 3 minutes (as opposed to 30 minutes). We scaled the threshold and payment accordingly.[17] The commitment value of the claw-back is reduced considerably in Treatment C because 3 minutes of bean sorting is hardly tedious and requires little commitment. We conjecture that maintaining self-control is easier for this task, and that leveraging loss aversion to increase effort levels is not optimal for most principals (except for the ones whose agents are extreme shirkers) as this would invite sub-optimally high levels of effort.[18] By comparing to what extent experienced participants self-

[17] We set T_3 by dividing the median productivity in the pilot phase by 10, and subjects were again informed they should do better than median performance in the pilot to qualify for the payment. The size of the payment was obtained by dividing the initial payment by 5; hence the threshold was set equal to 34 grams and the reward was now equal to UGS 500. Offering just UGS 250 (= UGS 2500 / 10) for sorting beans for 3 minutes (rather than 30) was deemed insufficiently salient, and hence we decided to offer a UGS 500 reward. Note this does not invalidate our comparisons as we increased the expected (per minute) payment for both the bonus and the claw-back regime in Task 2 of Treatment C. However, care should be taken when comparing productivity across treatment arms with different wages (something we will not do for the main analysis).

[18] In the language of the model in Section 2.1, for the three-minutes sorting task is a "low-α task." We assume that $\alpha_{ij} \approx 1$. In addition (but not captured by our model), one may argue that the 3-minutes bean sorting task is more "risky" because productivity may to a greater extent be beyond the control of the subject due to idiosyncratic shocks (e.g., sneezing, or an "unfavorable bag" of beans for sorting). This may also reduce the demand for a soft commitment device. Indeed, we find that the variance-to-mean ratio is greater for the 3-minutes task than the 30-minutes task, which is consistent with the idea of greater "riskiness."

select into either the bonus or the claw-back regime, we can test hypothesis 3. Theory predicts that entry in the claw-back should be greater in Treatment B than in Treatment C.[19]

Finally, Treatment D is designed to give some insights in the welfare effects of participating in a claw-back regime. We offered participants the choice between either participating in the claw-back regime, or accepting a fixed wage W_i, where i=1,2,3. We vary the fixed wage to allow construction of a demand curve for avoiding or self-selecting into the claw-back. We used three fixed wages: W_1=150, W_2=1200 and W_3=2400, and respectively 100, 200 and 100 participants were randomly allocated to one of these three sub-treatments.[20] While fixed wage earnings are unambiguous, there are two realizations of expected earnings in case the claw-back is selected, depending on assumptions about the information structure. First, subjects were informed that the threshold was placed at the median productivity level in the pilot study. Using this cut-off level as the threshold implies expected earnings in the claw-back regime equal to UGS 1250 (0.5 × UGS 2500). Second, subjects may have rational expectations about productivity and expect that productivity in the claw-back treatment will exceed productivity in the pilot study. Previewing our empirical results below, we find that no less than 60% of the subjects met the threshold in the claw-back regime. So subjects with rational expectations expect to earn 0.6×UGS 2500 = UGS 1500. We use both values in our welfare analysis below, where we compute the willingness to pay (WTP) for working in a claw-back regime for the median and mean subject in our sample. We compare (potential) utility losses from participation to increments in productivity – if any.

[19] One may be concerned that differences in choice for specific payment regimes between Treatments B and C is due to the size of the stakes (UGS 2500 in Treatment B, and UGS 500 in Treatment C). As discussed more fully below, however, we find that productivity (measured as the number of grams sorted per unit of time) in the low-stakes task was much higher than in the high-stakes task, which is at odds with the assumption that subjects "cared less." We believe subjects in both treatment arms considered their potential earnings as attractive.

[20] Putting relatively more mass on the middle value than for the extremes maximizes the power of detecting potential non-linearities in the demand function; see below for more information.

4. Results

For an overview of all experimental outcomes – productivity levels as well as regime choices – see Table A1 in Appendix A. To begin the results summary, we consider worker productivity in Task 1. As shown in Table 3, subjects in the claw-back regime (Treatment B) produced almost 25% more envelopes than those in the bonus regime (Treatment A). Our first main result thus confirms earlier findings, in other domains and by different "types" of participants (e.g., Hossain and List, 2012; Fryer et al., 2018; Levitt et al., 2012):

Result 1: *Average output is significantly higher if subjects are exogenously randomized into a claw-back regime (Treatment B) rather than into a bonus regime (Treatment A).*

Support for Result 1: The average number of envelopes folded is 20.57 in Treatment A, while it is 25.49 in the Treatment B. This difference is significant at $p < 0.0001$ according to a standard Mann-Whitney U-test.[21] ■

<< Insert Tables 3 and 4 about here >>

In line with hypothesis 1, Result 1 suggests manipulating reference points influences the average supply of effort, or that loss aversion can be leveraged to increase productivity – if subjects can exogenously be assigned to a claw-back regime.

We now probe into the propensity of subjects to select voluntarily into the claw-back regime, and consider how this propensity depends on previous experience with the incentive scheme (comparing regimes choices in Treatments A and B). Table 4 presents the shares of subjects choosing the claw-back regime for Task 2 for Treatments A, B, and C providing support for hypothesis 2:

[21] Table A1 shows that the number of envelopes folded is even higher in Treatments C and D than in Treatment B (although not significantly so). Comparing average productivity in Treatment A versus that in Treatments B-D yields a difference of 7.1 envelopes, and this difference is significant at $p < 0.000$.

Result 2: *Previous experience with the claw-back regime increases the propensity to choose the claw-back for Task 2. This is consistent with the interpretation that experience fosters "sophistication" and that sophisticated subjects acknowledge the commitment value of the claw-back.*

Support for Result 2: Of the 200 subjects in Treatment B (i.e., those exposed to the claw-back regime in Task 1), 81 chose the claw-back regime for Task 2, whereas only 46 subjects did so of the 200 subjects in Treatment A (having experienced the bonus regime in Task 1). This difference in shares (0.41 versus 0.23) is significant at $p = 0.0002$ according to the appropriate two-sided Equal Proportions test. ■

Recall that an additional interpretation exists for the data in Table 4, not based on experience fostering sophistication but on the propensity to "switch" to another incentive regime. For Treatment A, 154 people out of 200 remained with their original scheme, and in Treatment B "only" 81 stuck to what they had in the first round. Maybe switching behaviour is partly determined by inertia?

However, this does not seem to be the case. In Treatment C, with the non-tedious follow-up task of three minutes of bean sorting, subjects are not reluctant to switch to another management regime. No fewer than 144 (out of 200) now choose the bonus regime for Task 2, despite the fact that these subjects earlier worked under the claw-back regime. This is consistent with hypothesis 3 – selection into the claw-back is only sensible for tedious tasks requiring commitment.

Result 3: *Experienced subjects are less prone to select into the claw-back when confronted with a non-tedious 3 minutes task than with a tedious 30 minutes task.*

Support for Result 3: Focusing on those subjects who were exposed to the claw-back regime in Task 1, we find that 81/200 chose the claw-back regime when confronted with 30 minutes

of bean sorting in Task 2 (Treatment B), whereas only 56/200 did so when confronted with the 3 minute task (Treatment C). The shares in Treatments B and C (0.41 and 0.28) are significantly different ($p = 0.008$) according to a two-sided Equal Proportions test. Moreover, we find that there is no significant difference between the shares of subjects choosing the bonus regime if (i) Task 2 is not tedious (Treatment C) or (ii) subjects lacked previous experience with the claw-back (Treatment A). Inexperienced subjects are equally (un)likely to select into the claw-back regime for the heavy 30 minutes task as experienced subjects are for the light 3 minutes task (0.23 versus 0.28; $p = 0.251$ according to a two-sided Equal Proportions test).[22] ■

An interesting result is that productivity per unit of time is much higher in the 3-minutes task than in the 30-minutes task – both in the claw-back and the bonus treatment.[23] This is consistent with a convex (effort) cost curve. Interestingly, average productivity *is the same* for the claw-back and the bonus regimes in the 3-minutes task (56 versus 52 grams, $p = 0.363$). This insight is consistent with our theory in Section 2.1, if self-control varies across individuals as well as types of task (tedious or not). Non-random self-selection into the claw-back occurs, and workers who are not prone to shirking will avoid the claw-back (because the additional motivation would induce them to work too hard). For non-tedious tasks, the claw-back is only optimal for workers with the worst self-control problems, who need the claw-back as additional motivation to increase effort. Convergence of productivity levels across regimes may be the result if the most productive workers opt out of the claw-back.

[22] The reduced need for commitment is also evident when we compare productivity across incentive regimes in Treatment C. On average, only 8% more beans are sorted by subjects in the claw-back regime (56 versus 52 grams of beans sorted; see Table A1 in Appendix A), and this difference – due to the combination of both differences in incentives and selection – is not statistically significant ($p = 0.22$). Instead, in Treatment B we find that 18% more beans are sorted in the claw-back, and this represents a statistically significant difference. A similar result is found for the (absence of a) difference in the share of subjects reaching the threshold in Task 2, which is large and significant in Treatment B but not so in Treatment A; see Table A1.

[23] Comparing the productivity of those subjects who chose the bonus regime in Treatments B and C, the latter sorted 52 grams in the 3-minutes task, while the former only sorted, on average, 37 grams per 3 minutes in the 30-minutes task ($p<0.0001$). Similarly, comparing those who chose the claw-back regime in these two treatments, the ones who did so for the 3 minute task sorted 56 grams in the 3-minute task, while those who selected the claw-back for the 30 minute task did, on average, 40 grams in every 3 minute period ($p<0.0001$).

4.1 Robustness analysis: Education and sophistication

In the analysis above, we distinguish between experienced and inexperienced subjects, and we experimentally vary experience by random assignment to the claw-back in Task 1. It is possible that sophistication can also be fostered by other factors. Benjamin et al. (2013) report that people with higher cognitive abilities have lower levels of small-stakes risk aversion and short-run impatience. For example, they calculate that a one-standard-deviation increase in measured mathematical ability is associated with an increase of about 10 percentage points in the probability of behaving patiently over short-run trade-offs. In addition, they find that the same change in cognitive ability is associated with an increase of about 8 percentage points in the probability of behaving in a risk-neutral fashion over small stakes.

In this section, we use the cognition literature as a starting point to ask whether cognition, proxied by formal education, may also foster sophistication and facilitate recognition of the claw-back's potential as a commitment device. Since we cannot experimentally vary education levels, this analysis is based on non-experimental data, thus attribution rests upon additional assumptions. We regard these education results as a robustness analysis, and formulate the following hypothesis.

Hypothesis 2': *In the absence of personal experience with the claw-back regime, educated workers are more likely to select themselves into that regime than workers with little formal education.*

Our results are consistent with the literature, suggesting that cognition is correlated with our definition of sophistication. Specifically, for the sub-sample of inexperienced workers (i.e. subjects from Treatment A) we find that:

Result 2': *Inexperienced yet educated subjects are more likely to choose the claw-back regime for a tedious task.*

Support for Result 2': We define lower-educated subjects as those having received either no schooling, or just primary education. Using this definition, 118/200 subjects in Treatment A are coded as lower-educated. The percentage of lower-educated subjects choosing the claw-back regime for Task 2 is 16.1%, as opposed to 32.9% of the higher-educated subjects. This difference in shares is significant at p = 0.0054 according to the appropriate two-sided Equal Proportions test. ■

We perform the same test on experienced subjects from Treatment B and find that 46% of the more educated subjects self-selected into the claw-back regime for Task 2, compared to 36% of their less educated peers. This difference is not statistically significant at conventional levels ($p = 0.146$).

We also probe these issues in a regression framework. In Table 5, we present the results of a probit analysis in which we regressed the decision to self-select into the claw-back on previous success (meeting the payment threshold in the first task, or not) and an education dummy variable (columns (i)-(iii). We also explore whether the results are affected when including additional control variables (columns (iv)-(vi).

The regression analysis yields four results. First, previous success makes subjects more likely to self-select into the claw-back regime, especially if the previous task was completed under a claw-back regime and the task ahead is tedious (Treatment B). Second, subjects who did not complete secondary education are less likely to self-select into the claw-back treatment *unless* they had prior experience with the mechanism (compare Treatments A and B). Third, these results are unaffected when controlling for additional subject characteristics (compare the first two rows in columns (i)-(iii) to those in columns (iv)-(vi). Finally, the factors driving the selection decision in Treatments A and C are identical (qualitatively, and even quantitatively), but different from those in treatment B (except for prior success).

4.2 Towards an assessment of the welfare effects of claw-back regimes

Finally, we turn to the welfare implications of the claw-back. We use data collected in Treatment D to obtain measures for the relevant costs and benefits. In this treatment, all 400 subjects first participated in thirty minutes of producing envelopes. Their second task consisted of sorting beans for a period of 30 minutes (as in Treatment B), but they were offered either the choice between working (again) under a claw-back regime, *or for a fixed wage.*

When choosing the number of "fixed wages" for our welfare analysis we faced a trade-off. Increasing the number of fixed wage values generates information about more "intermediate points" on the aggregate demand curve for the claw-back, enabling more precise statements about welfare. But, this comes at the expense of statistical power. To test whether the demand function is linear or non-linear (either convex or concave), we follow the literature which argues that offering three wage rates maximizes the power of the statistical test—two at the extremes (close to the horizontal and vertical axes, each with 25% of the subjects) and one in the middle (with half of the subjects being offered that fixed wage; see McClelland (1995) and List et al. (2011). We implicitly assume the "demand function for the commitment device" is well-behaved in the sense that the share of subjects preferring the claw-back is a monotonously declining function of the fixed wage offered, and that the second derivative of this demand function is either (weakly) positive or (weakly) negative over the entire domain.

By implementing fixed wage rates at the extremes (close to 0 shillings, and close to 2500 shillings) we estimate the horizontal and vertical intercepts of the demand function, and the intermediate value of the fixed rate allows us to determine the second derivative of the demand function. If, when offered a fixed wage of 1200 shillings, the share of subjects preferring the claw-back would be (much) higher than 50%, then we learn that the demand

function for the claw-back is concave. Conversely, if the share is lower than 50%, then we learn that the function is convex.

We randomly assigned subjects to three fixed wage rates, UGS 2400 (100 subjects), UGS 1200 (200 subjects) and UGS 150 (100 subjects). The share of subjects choosing the claw-back decreases from 0.97 (for a wage of UGS 150) via 0.51 (UGS 1200) to 0.21 (UGS 2400). Hence, 51% of our subjects preferred the claw-back regime to receiving a fixed wage of 1200 shillings, meaning that the demand function for the claw-back is nearly linear. This linearity is a consequence of the distribution of behavioural preferences (self-control relative to loss aversion) in the population. The median value is UGS 1200 and – assuming a linear demand curve – the average is UGS 1396.[24]

So what fixed wage are experienced subjects willing to accept to avoid the claw-back when implementing Task 2 (30 minutes of bean sorting)? Based on our approximation of demand fitted through three fixed wages, we present the following result:

Result 4: *The median worker is (near-) indifferent between the claw-back contract and a fixed-wage contract paying its expected value, suggesting that overall welfare costs incurred by experienced subjects of being offered a take-it-or-leave-it claw-back contract are small.*

Support for Result 4: Recall that the expected value of participating in the claw-back equaled UGS 1250 (using a 50% threshold) or UGS 1500 (rational expectations); see Section 3. According to the demand curve fitted through the WTA data, the fixed wage that the median (average) subject is willing to accept to forego participating in a claw-back regime is UGS1200

[24] The average fixed wage at which our subjects are indifferent between working under the claw-back or the fixed wage regime is obtained by calculating the area underneath the "participation demand curve", which is obtained by regressing the percentage of subjects accepting the fixed wage on different fixed wages (for the full sample of 400 subjects in Treatment D).

(UGS 1396). The net welfare cost associated with (forced) participation in a claw-back regime is therefore negligible for our sample of experienced workers.[25] ■

More can be said about the welfare effects of introducing a *non-voluntary* claw-back contract for a sample of workers. While Result 4 suggests that the average welfare effect is negligible, this zero aggregate effect may hide considerable heterogeneity. Loss neutral workers are unaffected, but loss averse workers can become better or worse off. The theory in Section 2.1 is helpful in identifying winners and losers of claw-back regime assignment.

Figure 1 describes that, for any positive level of loss aversion, the relation between the planner-self's welfare in the claw-back and the (lack of) self-control is described by an inverted U-shaped curve. Workers who are not prone to shirking gain nothing from the claw-back, but are probably worse off because they are incentivized to supply too much effort. Workers who are extremely prone to shirking also gain little from the claw-back, as the incentive effect is not strong enough to meet the productivity threshold. Workers with intermediate self-control issues benefit from the claw-back and start supplying effort levels close to their optimal level. These workers have the largest willingness to pay for the claw-back as a commitment device in our experiment, and appear to choose voluntarily a "dominated contract" (as perceived through a non-behavioural lens). Non-shirkers and extreme-shirkers are likely better-off under a fixed wage regime.

A corollary of this result is that employers may be able to use the claw-back as a screening device to identify certain types of workers. When given the choice between a fixed

[25] Result 4 also suggests that the marginal cost of supplying effort must be low. This follows from the observation that despite the fact that the expected earnings are similar, those subjects who preferred to work under the fixed wage regime still supplied almost half of the effort put in by those who chose to work under the claw-back regime (on average 187 and 412 grams, respectively; see Appendix A, Table A1). It also follows from the simple observation that subjects in the fixed wage regime supply positive effort at all, despite the fact that this does not affect their earnings. One may conjecture that demand for the claw-back would be (even) greater among sophisticated subjects for a high-cost task inviting greater self-control challenges.

wage and the claw-back, workers self-select into the claw-back based on their propensity to shirk (and innate productivity, of course). "Weeding out" the extreme shirkers may affect the employer's pay-offs.[26] When we increase the fixed wage from 150 to 1200 and 2400, the share of subjects opting for the claw-back falls from 97% to 51% and on to 21%. However, the share of this (diminishing) subsample that actually meets the performance threshold increases: from 60% if (nearly) everyone participates to 74% when the top half chooses the claw-back, and to 81% when only the most confident 21% of subjects choose the claw-back. The incentive effect of the claw-back interacts with the changing composition of the worker sample, generating the result that productivity in the claw-back increases as the opportunity cost increases, but at a decreasing rate.

<< *Insert Table 6 about here* >>

Finally, it is possible to consider overall welfare – aggregating effects across "firms" and workers. Experienced workers who self-select into the claw-back regime produce more output for the firm, but the share of workers meeting the threshold to obtain or keep the UGS 2500 reward is higher too. About 78% of the workers choosing the claw-back met the payment threshold, compared to 49% of the workers choosing the bonus regime. This difference is significant at $p < 0.0001$, according to a two-sided Equal Proportions test. With self-selection, productivity is higher in the claw-back regime, but the average wage paid is also higher. As a result, the average cost per gram of beans sorted is UGS 0.21 for those who self-selected into the claw-back regime, compared to UGS 0.28 with endogenous selection into the bonus regime. Costs per unit of output are lower for the employer, so the employer is better off with the claw-back in our experiment. More work needs to be done to explore the generality of this insight.

[26] Following the discussion above, "weeding out" the extreme shirkers comes at the price of also losing the non-shirkers as both types are likely to opt for the fixed wage. In Section 2.1 we demonstrate this is not always the case: when employers offer the choice between the claw-back and a *low* fixed wage, then they can distinguish between low- and medium-shirkers on the one hand, and high-shirkers on the other hand.

5. Concluding remarks

Detailing the dark side of incentives has become an emerging point of research in the past decade. While certain types of incentive schemes have been shown to backfire, what has witnessed more limited attention is the potential deleterious effects of nudges, or subtle interventions meant to push individuals to conform to certain behavioural expectations. More generally, the welfare implications of nudges and commitment failures remain ill-understood. As governments around the world increasingly use behavioural manipulations to induce improved tax compliance (see, e.g., Hallsworth et al., 2014), and as market-based solutions to overcome commitment failures are taking off (e.g. Bryan et al., 2010), the stakes are heightened even further to deepen our understanding of the welfare effects of such interventions.

This study takes a step in that direction by linking a behavioral theory to a field experiment designed to measure potential negative consequences of leveraging loss aversion to motivate workers. Our results complement those obtained by Imas et al. (2016), and adds to them as we explicitly designed a test to identify the underlying mechanism of the choice for the claw-back regime – its potential usefulness as a commitment device for tedious or challenging tasks. Several interesting insights emerge, but perhaps the most important one is that the claw-back nudge does not, on average, have an adverse welfare effect on workers. This suggests that potential productivity gains observed from the direct incentives are not diminished through negative externalities of the incentive regime. Indeed, consistent with findings of Beshears et al. (2015) and Kaur et al. (2015) we find the opposite may be true: sophisticated workers learn to leverage loss aversion to become more productive. We also find tentative evidence that firms may be able to use the claw-back as a screening device – workers most and least prone to shirk will self-select out of the claw-back when given the choice. Screening therefore involves potentially complex tradeoffs for employers, and future work should explore this in more detail.

One natural question that arises is whether such effects can manifest themselves in the long run. We already discussed that experience fosters sophistication, so that the welfare effects of the claw-back evolve over time. But the salience of loss aversion may also be time-variant. Experience with the claw-back scheme might lead to less impact over time, as observed in trading markets where market experience attenuates loss aversion (List, 2003, 2004, 2011). However, these studies also show that extensive market experience is necessary to reduce the effects of loss aversion to zero. This represents a useful empirical exercise that future empirical researchers should tackle. For theorists, a full model describing how loss aversion evolves over time, and how its diminishment impacts the effects of nudges would be welcome. For practitioners, the results herein hold promise in that behavioural nudges can be used to motivate agents without being unraveled by the dismal side of the incentive.

Affiliations:

Erwin Bulte: Wageningen University, Development Economics Group, The Netherlands
John A. List: University of Chicago, Department of Economics, United States of America
Daan van Soest: Tilburg University, Department of Economics and TSC, The Netherlands

References

Abeler, J., Falk, A., Goette, L. and Huffman, D. (2011). 'Reference points and effort provision', *American Economic Review*, vol. 101(2), pp. 470-492.

Allcott, H. and Kessler, J. (2019). 'The welfare effects of nudges: a case study of energy use social comparisons', *American Economic Journal: Applied Economics*, vol. 11(1), pp. 236-276.

Augenblick, N., Niederle, M. and Sprenger, C. (2015). 'Working over time: dynamic inconsistency in real effort tasks', *Quarterly Journal of Economics*, vol. 130(3), pp. 1067-1115.

Benjamin, D.J., Brown, S.A. and Shapiro, J.M. (2013). 'Who is "behavioural"? Cognitive ability and anomalous preferences', *Journal of the European Economic Association,* vol. 11, pp. 1231–1255.

Beshears, J., Choi, J., Harris, C., Laibson, D., Madrian, B. and Sakong, J. (2015). 'Self-control and commitment: can decreasing the liquidity of a savings account increase deposits?', NBER Working paper 21474, Cambridge (MA): National Bureau of Economic Research.

Brink, A.G., and. Rankin, F.W. (2013). 'The effects of risk preference and loss aversion on individual behaviour under bonus, penalty, and combined contract frames', *Behavioural Research in Accounting,* vol. 25(2), pp. 145-170.

Bryan, G., Karlan, D. and Nelson, S. (2010). 'Commitment devices', *Annual Review of Economics,* vol. 2, pp. 671-698.

Camerer, C., Babcock, L., Loewenstein, G. and Thaler, R. (1997). 'Labor supply of New York city cabdrivers: one day at a time', *Quarterly Journal of Economics,* vol. 112(2), pp. 407-441.

Carroll, G.D., Choi, J., Laibson, D., Madrian, B. and Metrick, A. (2009). 'Optimal defaults and active decisions', *Quarterly Journal of Economics,* vol. 124(4), pp. 1639-1674.

Chetty, R., Looney A. and Kroft, K. (2009). 'Salience and taxation: theory and evidence', *American Economic Review,* vol. 99(4), pp. 1145-1177.

Cohn, A., Fehr, E. and Goette, L. (2017). 'Fair wages and effort provision: combining evidence from the lab and the field', *Management Science,* vol. 61(8), pp. 1777-1794.

Crawford, V. and Meng, J. (2011). 'New York city cabdrivers' labor supply revisited: reference-dependence preferences with rational expectations targets for hours and income', *American Economic Review,* vol. 101(5), pp. 1912-1932.

DellaVigna, S., (2009). 'Psychology and economics: evidence from the field', *Journal of Economic Literature,* vol. 47(2), pp. 315-372.

DellaVigna, S. and Malmendier, U. (2004). 'Contract design and self-control: theory and evidence', *Quarterly Journal of Economics,* vol. 119(2), pp. 353-402.

DellaVigna, S. and Malmendier, U. (2006). 'Paying not to go to the gym', *American Economic Review,* vol. 96(3), pp. 694-719.

DellaVigna, S., List, J. and Malmendier, U. (2012). 'Testing for altruism and social pressure in charitable giving', *Quarterly Journal of Economics*, vol. 127(1), pp. 1-56.

DellaVigna, S. and Pope, D. (2017). 'What motivates effort? Evidence and expert forecasts', *Review of Economic Studies*, vol. 85(2), pp. 1029–1069.

De Meza, D. and Webb, D. (2007). 'Incentive design under loss aversion', *Journal of the European Economic Association*, vol. 5(1), pp. 285-318.

De Quidt, J. (2018). 'Your loss is my gain: a recruitment experiment with framed incentives', *Journal of the European Economic Association*, vol. 16(2), pp. 522-559.

Engelmann, D. and Hollard, G. (2010). 'Reconsidering the effect of market experience on the "endowment effect"', *Econometrica*, vol. 78(6), pp. 2005-2019.

Fehr, E. and. Goette, L. (2007). 'Do workers work more when wages are high? Evidence from a randomized field experiment', *American Economic Review*, vol. 97(1), pp. 298-317.

Frederick, S., Loewenstein, G. and O'Donoghue, T. (2002). 'Time discounting and time preference: a critical review', *Journal of Economic Literature*, vol. 40(2), pp. 351- 401.

Fryer, R., Levitt, S., List, J.A. and Sadoff, S. (2018). 'Enhancing the efficacy of teacher incentives through loss aversion: a field experiment', Working Paper, Harvard University.

Fudenberg, D. and Levine, D. (2006). 'A dual-self model of impulse control', *American Economic Review*, vol. 96(5), pp. 1449-1476.

Genesove, D. and Mayer, C. (2001). 'Loss aversion and seller behaviour: evidence from the housing market', *Quarterly Journal of Economics*, vol. 116(4), pp. 1233-1260.

Goette, L., Fehr, E. and Huffman, D. (2004). 'Loss aversion and labor supply', *Journal of the European Economic Association*, vol. 2(2-3), pp. 216-228.

Gul, F. and. Pesendorfer, W. (2001). 'Temptation and self-control', *Econometrica*, vol. 69(6), pp. 1403-1435.

Gul, F. and Pesendorfer, W. (2004). 'Self-control and the theory of consumption', *Econometrica*, vol. 72(1), pp. 119-158.

Hallsworth, M., List, J.A., Metcalfe, R. Vlaev, I. (2017). 'The behaviouralist as tax collector: using natural field experiments to enhance tax compliance', *Journal of Public Economics*, vol. 148, pp. 14-31.

Hanemann, M., (1991). 'Willingness to Pay and Willingness to Accept: How much can they differ?', *American Economic Review*, vol. 81(3), pp. 635-647.

Hannan, R. L., Hoffman, V. B. and Moser, D. V. (2005). 'Bonus versus penalty: Does contract frame affect employee effort?', *Experimental Business Research*, vol. 2, pp. 151–169.

Harbaugh, W.T., Krause, K. and Vesterlund, L (2002). 'Risk attitudes of children and adults: choices over small and large probability gains and losses', *Experimental Economics*, vol. 5(1), pp. 53–84.

Henrich, J., Heine, S.J. and Norenzayan, A. (2010). 'The weirdest people in the world?', *Behavioural and Brain Sciences*, vol. 33, pp. 61–135.

Herweg, F. and Mierendorff, K. (2013). 'Uncertain demand, consumer loss aversion and flat-rate tariffs', *Journal of the European Economic Association*, vol. 11(2), pp. 399-432.

Herweg, F., Muller, D. and Weinschenk, P. (2010). 'Binary payment regimes: moral hazard and loss aversion', *American Economic Review*, vol. 100(5), pp. 2451-2477.

Hossain, T. and List, J.A. (2012). 'The Behaviouralist visits the factory: increasing productivity using simple framing manipulations', *Management Science*, vol. 58(12), pp. 2151-2167.

Imas, A., Sadoff, S. and. Samek, A (2016). 'Do people anticipate loss aversion?', *Management Science*, vol. 63(5), pp. 1271-1284.

Kahnemann, D., Knetsch, J. and Thaler, R. (1990). 'Experimental tests of the endowment effect and the Coase theorem', *Journal of Political Economy*, *vol.* 98(6), pp. 1325-1348.

Kaur, S., Kremer, M. and Mullainathan, S. (2015). 'Self-control at work', *Journal of Political Economy*, vol. 123(6), pp. 1227-1277.

Köszegi, B. (2014). 'Behavioural contract theory', *Journal of Economic Literature*, vol. 52(4), pp. 1075-1118.

Laibson, D. (1997). 'Golden eggs and hyperbolic discounting', *Quarterly Journal of Economics*, vol. 112, pp. 443-477.

Levitt, S., List, J.A., Neckermann, S. and Sadoff, S. (2016). 'The Behavioralist goes to school: leveraging behavioral economics to improve educational performance', *American Economic Journal: Economic Policy*, vol. 8(4), pp. 183-219.

List, J.A., (2013). ‘Does market experience eliminate market anomalies?’, *Quarterly Journal of Economics*, vol. 118(1), pp. 41-71.

List, J.A. (2004). ‘Neoclassical theory versus prospect theory: evidence from the marketplace’, *Econometrica*, vol. 72(2), pp. 615-625.

List, J.A., (2011). ‘Does market experience eliminate market anomalies? The case of exogenous market experience’, *American Economic Review Papers and Proceedings*, vol. 101(3), pp. 313-317.

List, J.A., Sadoff, S. and Wagner, M. (2011). ‘So you want to run an experiment, now what? Some simple rules of thumb for optimal experimental design’, *Experimental Economics*, vol. 14, pp. 439–457.

List, J.A. and Samek, A. (2015). ‘The Behaviouralist as dietician: leveraging behavioural economics to improve child food choice and consumption’, *Journal of Health Economics*, vol. 39, pp. 135-146.

Luft, J., (1994). ‘Bonus and penalty incentives contract choice by employees’, *Journal of Accounting and Economics*, vol. 18(2), pp. 181-206.

Mani, A., Mullainathan, S., Shafir, E. and Zhao, J. (2013). ‘Poverty impedes cognitive function’, *Science*, vol. 341(6149), pp. 976-980.

McClure, S., Laibson, D., Loewenstein, G. and Cohen, J. (2004). ‘Separate neural systems value immediate and delayed monetary rewards’, *Science*, vol. 306, pp. 503-507.

O’Donoghue, T. and Rabin, M. (1999). ‘Doing it now or later’, *American Economic Review*, vol. 89(1), pp. 103-124.

Rabin, M., (1998). ‘Psychology and economics’, *Journal of Economic Literature* vol. 36(1), pp. 11-46.

Samuelson, W. and Zeckhauser, R. (1988). ‘Status quo bias in decision-making’, *Journal of Risk and Uncertainty*, vol. 1(1), pp. 7-59.

Shiv, B. and Fedorikhin, A. (1999). ‘Heart and mind in conflict: the interplay of affect and cognition in consumer decision-making’, *Journal of Consumer Research*, vol. 26(3), pp. 278-292.

Thaler, R., (1980). ‘Toward a positive theory of consumer choice’, *Journal of Economic Behaviour & Organization*, vol. 1(1), pp. 39-60.

Thaler, R. and Shefrin, H. (1981). ‘An economic theory of self control’, *Journal of Political Economy* vol. 89(2), pp. 392-406.

Thaler, R. and Sunstein, C.R. (2008). ‘*Nudge: improving decisions about health, wealth, and happiness*’, New Haven (CT): Yale University Press.

Trope, Y. and. Fishbach, A. (2000). ‘Counteractive self-control in overcoming temptation’, *Journal of Personality and Social Psychology*, vol. 79(4), pp. 493-506.

APPENDIX A

A concise overview of all the results is presented in Table A1.

Table A1: Overview of performances and regime choices in Treatments A-D

Treatment A		**Treatment B**		**Treatment C**		**Treatment D**	
Task 1, producing envelopes							
bonus regime N = 200		claw-back regime N = 200		claw-back regime N = 200		claw-back regime N = 400	
Number of envelopes folded							
20.57 (6.71)		25.49 (8.55)		26.84 (8.61)		29.10 (8.97)	
Shares of subjects having met the threshold							
0.68		0.82		0.79		0.83	
Task 2, sorting beans							
Self-selection in one of following regimes		Self-selection in one of following regimes		Self-selection in one of following regimes		Self-selection in one of following regimes	
bonus regime N = 154	claw-back regime N=46	bonus regime N=119	claw-back regime N=81	bonus regime N=144	claw-back regime N=56	fixed wage regime N=178	claw-back regime N=221
Grams of beans sorted							
368.72 (108.18)	371.61 (93.28)	337.98 (119.16)	400.93 (86.78)	52.31 (18.73)	56.30 (23.61)	187.31 (93.99)	412.50 (104.60)
Shares of subjects having met the threshold							
0.63	0.56	0.49	0.78	0.88	0.80	NA	0.68

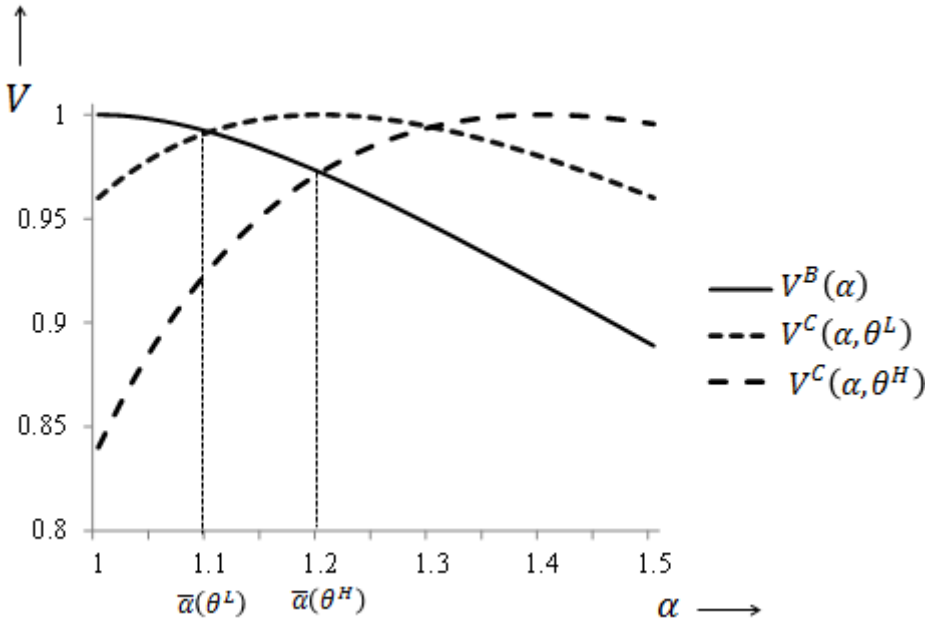

Figure 1: Pay-offs for the principals as a function of the lack of self-control parameter α under the two incentive regimes, and for two different degrees of loss aversion for the agent.

Table 1: Experimental design.

Treatment A		**Treatment B**		**Treatment C**		**Treatment D**	
Task 1, producing envelopes							
bonus regime N=200		claw-back regime N=200		claw-back regime N=200		claw-back regime N=400	
30 min envelope folding (A1) Receive payment if $Q_1 \geq T_1$		30 min envelope folding (B1) Return payment if $Q_1 < T_1$		30 min envelope folding (C1) Return payment if $Q_1 < T_1$		30 min envelope folding (D1) Return payment if $Q_1 < T_1$	
Task 2, sorting beans							
Self-selection in one of following regimes		Self-selection in one of following regimes		Self-selection in one of following regimes		Self-selection in one of following regimes	
bonus regime N=?	claw-back regime N=?	bonus regime N=?	claw-back regime N=?	bonus regime N=?	claw-back regime N=?	Fixed wage N=?	claw-back regime N=?
30 min bean sorting (A21) Receive payment if $Q_2 \geq T_2$	30 min bean sorting (A22) Return payment if $Q_2 < T_2$	30 min bean sorting (B21) Receive payment if $Q_2 \geq T_2$	30 min bean sorting (B22) Return payment if $Q_2 < T_2$	3 min bean sorting (C21) Receive payment if $Q_2 \geq T_3$	3 min bean sorting (C22) Return payment if $Q_2 < T_3$	30 min bean sorting (D21) Receive W∈{150, 1200,2400}	30 min bean sorting (D22) Return payment if $Q_2 < T_2$

Table 2: Summary statistics.

Variable	N	mean	median	sd
Tribe other than Musoga	1200	0.9075	1	0.289851
Education Level	1200	2.148333	2	1.272536
Gender	1200	0.550833	1	0.497617
Age	1200	34.72333	32	14.82143

Table 3: Number of envelopes made (Task 1) in Treatments A and B. [a]

		Number of envelopes produced
Incentive regime imposed in Task 1 (making envelopes)	Bonus regime (Treatment A)	20.57 (6.71) n = 200
	Claw-back regime (Treatment B)	25.49 (8.55) n = 200
		p < 0.000

[a] p-value obtained using a standard Mann-Whitney U-test.

Table 4: Propensity to choose the claw-back regimes in Treatments A-C.

	Share of subjects choosing the claw-back regime for Task 2 (sorting beans)	Differences in shares[a]
Treatment A	0.23 (46/200)	A-B: 0.18 p = 0.0002
Treatment B	0.41 (81/200)	B-C: 0.13 p = 0.0084
Treatment C	0.28 (56/200)	A-C: -0.05 p = 0.2513

[a] p-values obtained using a two-sided Equal Proportions test.

Table 5: Probit regression results of the decision to choose the claw-back regime in treatments A-C.

	(i)	(ii)	(iii)	(iv)	(v)	(vi)
Treatment	A	B	C	A	B	C
Threshold Task 1 met (Y/N)	0.453*	0.706***	0.416+	0.474**	0.724***	0.419+
	(0.239)	(0.262)	(0.257)	(0.238)	(0.279)	(0.255)
Secondary education or higher	0.459**	0.201	0.335*	0.465**	0.0741	0.384*
	(0.206)	(0.185)	(0.196)	(0.207)	(0.201)	(0.204)
Tribe other than Musoga				0.109	0.631*	0.0213
				(0.356)	(0.365)	(0.507)
Age				0.000587	-0.0145**	0.00125
				(0.00789)	(0.00600)	(0.00764)
Female				0.0863	-0.373*	0.212
				(0.206)	(0.194)	(0.195)
Constant	-1.279***	-0.919***	-1.061***	-1.468***	-0.788+	-1.280**
	(0.214)	(0.249)	(0.235)	(0.483)	(0.538)	(0.618)
N	200	200	200	200	200	200
Wald Chi2	10.67	9.07	6.65	12.05	20.47	8.93