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

Erwin Bulte[a], John A. List[b] and Daan van Soest[c]

[a]Wageningen University, Development Economics Group, The Netherlands
[b]University of Chicago, Department of Economics, United States of America
[c]Tilburg University, Department of Economics and TSC, The Netherlands

Abstract: Social scientists have recently explored subtle approaches ("nudges") to encourage certain behaviors across a variety of situations, from energy conservation to healthy dieting. Focusing on how framing of gains and losses affects productivity, we take a step toward understanding the power and welfare implications of nudges by conducting a field experiment in peri-urban Uganda. Comparing 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. When subsequently testing whether workers prefer the gain or loss contract, we obtain tentative evidence that the welfare costs of using the loss contract are quite modest. Indeed, some workers value the loss contract more than the gains contract, perhaps due to its commitment features. Beyond providing a first step towards a roadmap for examining welfare effects of other behavioral interventions, our study suggests that one important nudge in the workplace does not come with deleterious effects.

Keywords: nudge, reference-dependent preferences, effort, labor productivity, field experiment, commitment, claw-back.

JEL Code: C93, D03, J3.

* We want to thank two anonymous reviewers for constructive and insightful comments on the paper. Remaining mistakes are our own. We thank John Sseruyange for his excellent research assistance when implementing the experiments in the field. List thanks the Sloan Foundation for financial support.

1. Introduction

Constructing incentive schemes to motivate individuals to take a particular course of action is perhaps among mankind's oldest activities. Whether at home, at school, in the office, or strolling in public quarters, we are surrounded by incentives created to induce a particular set of behaviors. For their part, economists have produced a rich assortment of models to help us understand situations in which the various elements of incentive contracts should enforce productive behaviors, and when they might exacerbate misconduct. As an important complement to standard models, the field of behavioral economics is evolving rapidly to explain departures from rational models. Early work focused on lab experiments documenting deviations from "rational behavior" – *Homo economicus* style – and subsequently attempted to capture "behavioral 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 behavioral economics more seriously when constructing incentive contracts. While some analyses focus on so-called exploitative contracts, where firms seek to take advantage of behavioral "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 behavioral 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 behavioral 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 the 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

[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)).

evidence in the affirmative. They exogenously shift agents' expectations about the remuneration they may receive, and find that effort is adjusted 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. The efficacy of such a claw-back (or penalty) incentive scheme was first tested in a field experiment due to Hossain and List (2012), who showed that loss aversion can be leveraged to increase labor productivity. Collaborating with a Chinese electronics company, they implement a simple framing experiment where a random subsample of workers is provisionally promised a bonus, to be paid at the end of the study period, but the salary enhancement is reduced 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. (2012) 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, leveraging loss aversion via a penalty regime is found to be effective. Similar results are obtained when incentivizing students, rather than teachers (Levitt et al. (2012)). The claw-back incentive regime again outperforms a conventional bonus regime, awarding good performance on tests ex post.[2]

While this literature focuses on how loss aversion promotes effort for exogenously-imposed incentive regimes, a small number of very recent studies considers the demand for incentive regimes – allowing for a fuller consideration of potential welfare effects. One major question is whether agents may 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 sophisticated (aware of their self-control issues), 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. This is consistent with evidence provided by Kaur et al. (2015) who demonstrate that, on average, Indian workers voluntarily set positive targets in a piece rate

[2] Loss aversion does not always appears 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 under-researched, and important, line of work.

regime penalizing them for falling short of that target (by cutting the piece rate). Similarly, Beshears et al. (2015) argue that agents value commitment by studying savings behavior. They compare deposits in so-called commitment and liquid accounts, or accounts with and without limitations on withdrawals. They find positive demand for commitment accounts, even in the absence of any interest premium (see also Ashraf et al. 2006). Moreover, when the commitment and liquid account have the same interest rate, demand for the commitment account increases with the degree of illiquidity. It appears as if, in their current decisions, many agents are willing to incentivize or restrict their own future behavior.

The commitment value of dominated contracts has straightforward implications for the case of bonus versus penalty regimes. Loss averse agents should prefer bonus contracts over penalty contracts when the underlying payoffs are identical. Earlier work by, 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 this mechanism explains why agents prefer penalty regimes over bonus regimes in their lab experiment. De Quidt (2017) also finds that job offer acceptance rates were higher if the job description was framed as a penalty rather than a bonus contract. However, and surprisingly, agents in his study also preferred the penalty contract for a coin-tossing task where commitment is unimportant. De Quidt (2017) therefore argues against self-commitment as the main explanation, and instead proposes that the salience of the payment structure matters. Even if bonus and penalty regimes are isomorphic, they may appear different from the perspective of the worker. Survey evidence suggests workers may regard the penalty contract as "better paid."

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.[3] Our field experiment mimics the workplace and is set in a novel context (peri-urban Uganda), introducing a large sample of "non-standard" subjects (i.e. non-WEIRD respondents; Henrich et al. 2010). Second, as do Imas et al. (2016) and De Quidt (2017), we allow subjects to choose between the two regimes in a

[3] For example, DellaVigna and Pope (2016) 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.

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, fostering appreciation of the regime's value as a self-commitment device. Previewing our results, we find the share of subjects choosing the claw-back regime for a specific task is higher among subjects who have previously been exposed to that regime. While we view this result as evidence that previous experience increases demand for the claw-back as a commitment device, alternative explanations may be valid too – including the quite mundane hypothesis that preferences for the regime are simply "sticky". 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.[4] This simple model, sketched in Appendix 1, predicts that demand for the claw-back as a soft commitment device should be smaller the less tedious (or challenging) the task is. We thus experimentally vary how tedious the task is, and find that experienced subjects are less likely to select into the claw-back if the task is less tedious. This outcome is consistent with the claw-back contract being a soft commitment device. To the best of the knowledge, we are the first to explicitly build the "varying demand for soft commitment devices" into our experimental design, implying our results not just complement but also provide additional insights beyond findings documented in companion papers like De Quidt (2016) and Imas et al. (2016). Third, we take a small step in the direction of understanding the welfare implications of nudges, using the claw-back as our case study.[5] While manipulating incentive regimes in accordance with insights from behavioral economics may increase productivity, the overall welfare effects for agents are unclear theoretically. We measure the (average and mean) utility loss associated with the claw-back, and contrast this loss to the productivity gain.

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 acts in accord with a model of a sophisticated agent, who has "learned" how to use

[4] 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 fulfil a plan for future behaviour that would otherwise be difficult owing to intrapersonal conflict stemming from, for example, a lack of self-control."

[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 2015).

the claw-back to commit to supplying higher effort levels in a subsequent task. This evidence of learning is consistent with Kaur et al. (2015), who also find that demand for dominated contracts as a commitment device increases as agents gain experience with them. Such commitment is not optimal for all workers, however, depending on the combination of behavioral biases affecting the worker's behavior, and the level of sophistication. Loss aversion and sophistication vary across agents (e.g. Beshears et al. 2015). Commitment is also unimportant for certain tasks, such as De Quidt's coin toss. Supporting our interpretation, we find little demand for the claw-back, even among sophisticated workers, when the task at hand is not very challenging. Third, we demonstrate that some workers self-select out of the claw-back regime, even at a financial cost. For example, inexperienced workers tend to avoid the claw-back contract, consonant with earlier field experiments showing that inexperienced agents have preferences that exhibit considerable loss aversion (e.g., List (2003), (2004), Englemann and Hollard (2010)).

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. Accordingly, it does not appear that any productivity gains experienced by the firm are diminished due to worker utility losses. This result provides an initial indication of the potential efficacy of using behavioral insights, such as loss aversion, to encourage workers to put forth higher effort levels. We also document quite a bit of heterogeneity in terms of welfare effects for our subjects, and argue this can perhaps be exploited to use claw-back regimes as a screening device.

The remainder of our study is organized as follows. In Section 2 we discuss the workings of the claw-back regime, and define concepts such as sophistication and commitment in our context. A formalization of our ideas is presented in Appendix 1. In Section 3, we introduce our field experiment, describe our data, and outline our identification strategy. In Section 4, we present our 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 behavioral economics that have prominently found their way in 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 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.

While we do not aim to develop a full-fledged model of evolving sophistication and the implied consequences for productivity and self-selection into incentive regimes, we wish to sketch a simple model that guides interpretation of the empirical results, keeping in mind Köszegi's (2015) insights (see Appendix 1 for a formalization of these ideas).

We assume that workers suffer from both loss aversion and a self-control problem – possibly to varying degrees. 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)). *Ex ante*, a worker prefers to exert high effort and work hard. But when the time to work arises he is tempted to procrastinate or shirk. Bryan et al. (2010) refer to this as outcomes where the beliefs of agents regarding costs and benefits of specific activities differ over time. 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.[6] 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."[7]

Assume each worker consists of two personalities: the rational and reflexive principal ("the planner"), aiming to maximize overall utility, as well as a biased agent ("the implementer"). The decision problem we study may be regarded as a game between a biased "self," responsible for deciding about effort levels, and a rational "self," responsible for selecting the incentive regime in which the former works. The underlying idea is that the

[6] See Thaler and Shefrin (1981) for a pioneering contribution focusing on the consumption-savings problem, and Fudenberg and Levine (2007) for a more general treatment.

[7] Fudenberg and Levine (2007) 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).

rational principal might try to manipulate the agent's behavior by choosing a certain incentive structure.

Our line of arguing is inspired by pioneering work of O'Donoghue and Rabin (1999), but deviates from that model in various respects. They assume a game between a principal and agent, where the agent "suffers" from one behavioral bias – a self-control problem. The agent may for example overestimate 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 or not she "knows" the utility function of the agent. In contrast, we assume (i) the agent suffers from *two* behavioral biases (self-control as well as loss aversion), and (ii) all principals know their agents' utility function. Within our experimental setting—adult workers in a naturally-occurring setting—it seems realistic to assume that (through introspection) the principals embodied in these individuals have had ample opportunity to learn about the true nature of their agents. For example, people have learned whether they are prone to procrastinate. Hence, our theoretical definition of sophistication does not hinge upon knowledge of the agent's utility function. Instead, we adopt a cognitive perspective, and define principals as "sophisticated" if they understand that one behavioral anomaly can be leveraged to tackle the other. We allow experience in the experiment and education to foster sophistication, and will subsequently test whether experience and education actually matter in the empirical analysis.

Consider the case where a worker performing a task has to choose between being governed by a bonus or claw-back regime. Across regimes, the financial rewards are the same and only depend on whether or not the worker meets a performance threshold. The regimes differ in the timing of provision of the reward. In the bonus regime, the worker receives the reward after meeting the performance threshold; in the claw-back regime the reward is paid up-front, but the worker must return it in case she fails to meet the threshold (see Fryer et al. (2012) and Levitt et al. (2012)). For workers without reference-dependent preferences this distinction is immaterial. However, workers with loss-averse "implementers" will experience a loss in utility when they have to return the payment in case they fail to meet the threshold under a claw-back regime. Moreover, this loss is larger than the increase in utility when receiving the equivalent payment in the bonus regime (if the performance threshold is met). Since the claw-back has psychological consequences above and beyond the financial consequences of not receiving the payment, Bryan et al. (2010) refer to such devices as "soft commitment."

The intuition for how loss aversion might influence behavior in this setting is that since all principals know their agents' utility function, they recognize their agents might fail to supply sufficient effort (due to the self-control problem). Further, they are aware of the potential utility loss invited by a failure to meet the threshold (the loss aversion problem). Anticipating this potential utility loss, a non-sophisticated principal is perhaps prone to choosing the bonus regime. But sophisticated principals anticipate that the claw-back regime may discipline the loss-averse agent, inducing him to work harder and increasing the odds of actually earning the reward. Loss aversion is leveraged to overcome the self-control problem, and the claw-back regime acts as a commitment device for the sophisticated principal. However, it can be shown that even sophisticated principals may prefer the bonus regime – if their agent is sufficiently loss averse, so that the utility loss associated with the penalty looms particularly large, or if (the lack of) self-control is not likely to affect effort too much, as in a coin tossing task. The precise relationship between the principal's welfare and the agent's characteristics and performance in the two regimes are derived in Appendix 1.

Based on this reasoning, we now present our hypotheses. When agents are randomized into regimes, self-control and loss aversion are orthogonal to treatment status, and distributions of these behavioral traits should be identical across regimes. We expect that a non-negligible share of our subject pool is loss averse, and works extra hard to avoid the penalty in a claw-back regime. Hence:

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 two types of principals. The first type of principals is "not sophisticated." While cognizant of their agents' behavioral biases (self-control and loss aversion), such principals lack the insight that the choice of the reference point affects effort supply by the agent, and fail to realize that the probability of achieving the threshold is higher in the claw-back regime. Assuming that in case of indifference the principal chooses the regime that maximizes the agent's welfare, non-sophisticated principals weakly prefer to select their agents into the bonus regime – the non-sophisticated principal expects the probability of reaching the threshold is the same in both regimes, while the agent's disutility of not reaching the threshold is smaller in a bonus regime than in the claw-back regime. In contrast, a sophisticated principal recognizes that loss aversion can be leveraged to reduce shirking, and may prefer his agent to work in a claw-back regime

unless this would induce the agent to work too hard – if the agent's lack of self-control is small relative to his level of loss aversion. In the empirical analysis we use (experimentally induced differences in) experience with the claw-back as a proxy for sophistication, reflecting that many subjects quickly update their preferences for commitment devices after "experiencing" the consequences of their behavioural anomalies (Augenblick et al. 2015).

We can now specify the following hypotheses.

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

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.*

We seek to test hypotheses 2 and 3 by experimentally varying the level of experience as well as by varying the extent to which lack of self-control is likely to be important for the task ahead (i.e. the nature of the task – demanding, or not).

Next, turn to comparative statics with respect to the level of loss aversion. While we cannot exogenously vary loss aversion levels of our respondents, our observational data allow us to probe some further predictions of our theory. This part of the analysis necessarily is more tentative, and identification rests upon additional assumptions. Observe that to non-experienced workers the possibility of losing their reward is the most salient feature of the claw-back regime, and hence they are less likely to select into the claw-back regime if they are more loss averse. The situation is more complex for experienced principals. For "light tasks" that require little commitment, choosing the claw-back is only optimal for workers who are not very loss averse. In contrast, for "heavy tasks" that require commitment and discipline, an experienced principal might prefer the claw-back – even if her agent is quite loss averse. We state the following hypotheses regarding how loss aversion affects the regime choice by experienced and non-experienced workers:

Hypothesis 4*: (4a) For higher levels of loss aversion, non-experienced workers are less likely to choose the claw-back regime.*

(4b) For higher levels of loss aversion, experienced workers are only less likely to choose the claw-back when the task is light ("non-tedious") and does not require discipline.

(4c) The propensity to choose the claw-back does not vary significantly with the level of loss aversion if (i) subjects are experienced and (ii) the task is tedious and requires discipline.

Finally, we take a first step towards considering the welfare effects of the two incentive regimes. The claw-back introduces a direct negative effect, caused by the fact that some workers may fail to meet the threshold, and have to return their reward. Fearing such outcomes, agents may also work "too hard" – supplying too much effort (see Appendix 1). Such workers, if they are loss averse, will suffer a utility loss. Alternatively, a subsample of (experienced) workers realizes the scope for exploiting the mechanism as a commitment device, and as a result will reach outcomes that are closer to the principal's first best. Such principals are better off and, when given the choice, may be willing to pay a positive amount to self-select into the claw-back regime. The net welfare effect for a group of workers will depend on the magnitude of these effects, and on the population share of experienced principals in the population.

Hypothesis 5*: The net welfare effect of exogenously introducing a claw-back incentive regime is not unambiguously negative.*

Note that we framed our main hypotheses in terms of experience with the claw-back, as the theory (implicitly) assumes that agents have a good understanding of what a bonus scheme entails (but not necessarily a similar level of understanding of what the claw-back regime entails). Indeed, in the region in which we implemented our field experiment, peri-urban Kampala, many people have experience with ex-post bonus payments that are conditional on successfully implementing contracted tasks. About 85% of our subject pool is employed in the informal sector, where payments are often conditional on the achievement of specific targets. While bonus payments are rarer in the formal sector in Uganda, they are part of the remuneration of about a quarter of those participants with formal employment (according to our data). As such, if experience affects sophistication, it is especially by offering subjects experience with a claw-back payment scheme compared to those who we do not offer such experience.

Education may also matter, however, so we introduce a "sophistication production function" to our model in the appendix. The arguments in this sophistication production function include experience and education, and we treat sophistication as a latent variable that is especially relevant for our theory.

3. Experimental design and data

To test our theoretical predictions and to probe the scope for leveraging loss aversion to enhance labor productivity, 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-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 tasks, j to denote subjects, and z to denote treatments. We selected folding envelopes and sorting beans because output Q_{ijz} (i=1,2; j=1..,1000; 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, enabling 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 a fixed wage and piece rate compensation 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 our two main treatments (A and B) 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.[8]

Next, turn to our two-task experiment, summarized in Table 1. Our 1000 individuals received a show-up fee of UGS 2000.[9] Since we experimentally vary the level of experience

[8] Subjects are 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)).

[9] 1 USD = 2800 UGS, and the average daily wage was about 6000 UGS or USD 2.14.

(as a proxy for sophistication) of the principals, and because personal experience may invite steep learning effects, we seek to vary the level of sophistication by *exogenously* allocating 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 a bonus regime, Treatment A, in which workers received a payment of UGS 2500 upon completing 18 envelopes, or more. Before Task 1 all subjects were informed about "their" payment regime in the first task. For Treatment A, in which Task 1 was to be performed under a bonus regime, this implied 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 800 subjects who were randomized into Treatments B-D, in which the claw-back regime was in place for Task 1, also received the same information about the task at hand, but we explicitly invited them to also 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). They were allowed to keep the payment after the task conditional upon completing 18 or more envelopes. If they failed to fold 18 envelopes within the thirty minute time period, they had to return the envelope with 2500 UGS to one of the experimenters. Comparing output across the treatment arms allows us to directly 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 impact of loss aversion on effort. In contrast, workers in the other three treatments (B-D) received first-hand experience with the claw-back in the first task, and may have become sophisticated.[10]

After the first task, participants were given a short break. During the break each of the participants was informed about whether they had met the threshold for Task 1. Those who had reached the threshold in Treatment A received their payment envelope with UGS 2500, and those who failed to reach the threshold in Treatments B-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 any concerns that having the payment envelope on one's

[10] 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) 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.

desk rather than at the experimenter's desk may have resulted in differential levels of trust in that the payment would be forthcoming in Treatments B-D as opposed to Treatment A (despite the fact that all subjects had been given the opportunity to verify the content of the payment envelopes).

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 were in place for their Task 1 (the bonus regime for Treatment A and the claw-back regime for Treatments B-D), and they were also explained the details of the other regime as outlined above. That is, subjects in Treatment A received additional information on the functioning of the claw-back regime, whereas subjects in Treatments B-D were informed of the details of the bonus regime.

Upon completion of the instructions, subjects were invited to select the incentive regime they preferred for completing the second task. Hence, they were *not* randomly allocated to any incentive regime, but were *free to self-select* into either the bonus or claw-back regime. Details of the task, as well as of both regimes, were spelled out to all participants. Task 2 in Treatments A and B lasted 30 minutes, presumably making bean sorting quite tedious (so that some discipline is required). Comparing self-selection of subjects in Treatments A and B, we can test hypothesis 2, whether experienced participants are more likely to choose the claw-back regime. While our preferred channel linking Task 1 and the choice for specific payment regimes in Task 2 is differences 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).

We also experimentally vary the extent to which self-control affects decisions and outcomes. For this purpose we implement Treatment C, where (experienced) subjects are also offered the choice between the claw-back and bonus regime for Task 2, but where the second task only lasts 3 minutes (as opposed to 30 minutes). We scaled the threshold and payment.[11]

[11] 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)

The commitment value of the claw-back is reduced considerably, as 3 minutes of sorting is not tedious and does not require much discipline.[12] In addition (but not captured by the model in Appendix 1), 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, an "unfavorable bag" of beans for sorting). 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." For these two reasons – the 3-minutes task is less tedious and subjects have reduced control over their own productivity[13] – demand for commitment should go down. By comparing how experienced participants (i.e. from treatment arms B and C) self-select into bonus or claw-back regimes, this treatment allows us to test hypothesis 3. Theory predicts that experienced participants will only self-select into the claw-back when the "principal" seeks to strategically tie "his agent" to the mast. That is; entry in the claw-back should be greater in treatment arm B (30 minutes of sorting) than in treatment C (3 minutes).[14]

Next, Treatment D was 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 Y_i, where i=1,2,3. We varied the fixed wage to allow construction of a demand curve for avoiding or self-selecting into the claw-back. We used three fixed wages: Y_1=150, Y_2=1200 and Y_3=2400, and respectively 100, 200 and 100 participants were randomly allocated to one of these three sub-treatments. While fixed wage earnings are unambiguous, there are two realizations of expected earnings in case

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 obviously be taken when comparing productivity across treatment arms with different wages (something we will not do for the main analysis).

[12] The literature provides some guidance into the relevant parameter space. In particular, Trope and Fishbach (2000) use an experimental design to analyze self-imposed penalties in the context of a costly task, and found that subjects confronted with a more difficult task on average choose larger penalties. Bryan et al. (2010, p.680) conclude that "agents seemed to demand commitment to help them with a difficult task..." Hence, when the task at hand is not tedious and does not require additional encouragement or discipline, we should expect that none of the workers should face self-control issues. For such non-tedious, commitment is not important and the claw-back only introduces (potential) costs in the form of loss aversion and over-supply of effort. For demanding or tedious tasks, such as 30 minutes of bean sorting, commitment may matter. This issue is empirically tested below.

[13] If effort is relatively less important, the value of the commitment device decreases, and loss averse principals are more likely to select their agents into the bonus contract.

[14] 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, and average productivity per unit of time being higher in both the bonus and in the claw-back regimes indicates that there was less demand for a commitment device in the light task.

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 fully 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.[15]

<< Insert Table 1 about here >>

Finally, we conjecture that the behavior of subjects in the experimental tasks is correlated with their aversion to losses. To collect data on this variable, and others, we invited subjects to participate in a small entry survey. We asked them the usual set of questions (age, tribe membership, religion, status in the household, marital status, education levels), but also asked hypothetical questions to probe levels of risk and loss aversion. Specifically, to obtain a metric for loss aversion, we asked all participants to answer the following survey question:

> "*Here is a […] hypothetical example. Assume you receive UGS 5000. Next, you have the choice between participating or not participating in a lottery.*
>
> *In the lottery, a coin is flipped. If it comes up heads, you need to pay 2000 from the 5000 you just received, and you would go home with 3000. If it comes up tails, you can keep your 5000.*
>
> *You can also decide to not participate in this lottery, but instead pay a fixed amount.*
>
> *What is the largest amount, of the 5000, you would be willing to pay so that you do not have to participate in the lottery?*

[15] Observe that respondents do not choose between the bonus and the fixed wage. This means that we cannot determine whether adoption rates are driven by loss aversion or risk aversion. To probe this issue somewhat, we estimate a Probit model regressing preference for the claw back on risk aversion and other co-variates such as age, education, performance in the first task. We consistently find that risk aversion does *not* explain choice of payment regime. Details are available on request.

Again, remember, in the lottery there is a 50% chance of losing 2000, and a 50% of losing nothing."

Loss-averse subjects should be more prone to take a gamble (with a 50% chance of *not* losing any money, if the coin should come up tails) than to incur a sure loss (by paying an amount of money with certainty). Hence, we construct a variable *LossAversion*, defined as the expected value of the loss, 1000, minus the subject's answer to the above question, and subsequently divide this number by 1000. Higher values of *LossAversion* indicate the subject is more loss averse.[16] As a robustness test, we have also used our observational data to construct another measure for loss aversion, namely whether or not the subject's utility function is "kinked" at the origin. For this purpose, we also need a measure of the certainty equivalent in the gains domain (or a measure of risk aversion). Hence, we asked a similar question as the one above to measure risk aversion. Our kink variable captures whether the slopes of the utility function in the gains and loss domain (based on certainty equivalents) are the same.

Note that we did not financially incentivize the loss aversion and risk aversion questions; this is because we only rely on a metric to *rank* subjects in accordance with their preferences – not to exactly measure each subject's level of preference. There is no evidence of which we are aware that hypothetical bias will reverse our ordinal ranking. If the hypothetical nature of the exercise introduces noise, this should cause us to reject fewer nulls with the data exercise. Descriptive statistics of our main dependent and explanatory variables are summarized in Table 2.

<< Insert Table 2 about here >>

The average level of loss aversion is negative – subjects are willing to pay more than the expected value of the loss to avoid participating in the gamble. As indicated by the standard deviation, there is substantial heterogeneity in our subject pool, and the median participant is loss neutral. About fifty percent of the participants in our sample display loss aversion, and any impact of the framing treatments should come from this sub-sample. As expected, the average

[16] Loss aversion predicts that the utility function is convex in the loss domain, or that the willingness to pay to avoid losing money in a lottery is smaller than the expected value of the loss. We measure this as follows: the *smaller* the amount one is willing to give up for certain to avoid participating in a risky gamble with a large potential loss, the more loss averse the individual is. In other words, more loss averse subjects display smaller certainty-equivalent payments. Fehr and Goette (2007) classify subjects more precisely by having them make two participation decisions: one for a lottery with positive expected value (and a chance of losing money), and one with the same lottery implemented multiple times. Viewing the complexity of our experimental design and the type of subjects participating in the experiments, we decided to present the simplest possible gamble in the loss domain; as in Harbaugh et al. (2002).

metric of risk aversion is positive, and the median participant completed secondary education: values of schooling above (below) 2 imply that the participant received more (less) than secondary schooling. Slightly more than half of our participants are female, and the tribal affiliation is quite diverse (with less than 10% of the participants belonging to the tribe represented most).

4. Results

To begin our summary of the experimental data, we consider the aggregate data across treatment arms, and do not impose parametric assumptions. For an overview of all experimental outcomes – productivity levels as well as regime choices – see Table A1 in Appendix 2. Consider first 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. (2012), 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.[17] ■

<< Insert Tables 3 and 4 about here >>

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

We now probe into the propensity of subjects to voluntarily select 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

[17] 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 $p < 0.000$.

subjects choosing the claw-back regime for Task 2 for Treatments A-C, providing support for hypothesis 2:

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 the propensity to switch is partly determined by inertia. However, this does not seem to be the case. In Treatment C, identical to Treatment B except that Task 2 is less tedious, subjects are not reluctant to switch to another management regime. No fewer than 144 (out of 200) choose the bonus regime for Task 2, despite the fact that these subjects all completed Task 1 under the claw-back regime. This is at odds with the conjecture that inertia induces people to "stick" with the regime they were randomized into for the first task, while it 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 light 3 minute task than with a 30 minute 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 (0.405 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 minute task as experienced subjects are for the light 3 minute task (0.23 versus 0.28; p = 0.2513 according to a two-sided Equal Proportions test).[18] ■

The assumption that the 3-minutes task does not require commitment is supported by the data. Not only is productivity per unit of time much higher in the 3-minutes task than in the 30-minutes task – both in the claw-back and the bonus treatment[19] – we also find that productivity *is the same* for the claw-back and the bonus regimes in the 3-minute task (56 versus 52 grams, p = 0.363). This suggests that the great majority of subjects can work hard for 3 minutes, and do not need commitment for that task.

We now introduce our survey-based measure of loss aversion, which allows us to consider hypotheses 4a-c on how loss aversion mediates the regime choices in Treatments A-C. We first examine how the degree of loss aversion affects self-selection for inexperienced participants (Treatment A). Consistent with our predictions we find:

Result 4a: *Inexperienced subjects are less likely to choose the claw-back regime if they are more loss-averse.*

Support for Result 4a: In Treatment A the average degree of loss aversion of subjects choosing the bonus regime is significantly higher than that of subjects choosing the claw-back regime (*p* = 0.006 according to a two-sided *t*-test). ■

Next, we ask how loss aversion affects self-selection of experienced subjects. When the commitment value is unimportant (i.e., when the task is "light"), experienced subjects should avoid the claw-back. Indeed, this prediction is supported by the data:

Result 4b: *If self-control is not important (i.e. for "light tasks"), experienced subjects are less likely to choose the claw-back regime if they are more loss-averse.*

[18] 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), 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.

[19] For example, in the bonus regime subjects sort 52 grams per 3 minutes in the 3-minutes task, and only 34 grams per 3 minutes in the 30-minutes task (p<0.01).

Support for Result 4b: The average degree of loss aversion of subjects choosing the bonus regime is significantly higher than that of subjects choosing the claw-back regime in Treatment C ($p = 0.062$ according to a two-sided *t*-test). ■

Outcomes are different for "heavy tasks", where potential utility losses due to loss aversion and over-supply of effort may be offset by the commitment value of the claw-back:

Result 4c: *Loss aversion does not affect the propensity to choose the claw-back regime if (i) subjects are experienced* and *(ii) the task is heavy.*

Support for Result 4c: The average degree of loss aversion of subjects choosing the bonus regime is not significantly smaller than that of subjects choosing the claw-back regime in Treatment B ($p = 0.452$ according to a two-sided *t*-test). ■

Additional support for Results 4a-c comes from a parametric analysis. Table 5 presents the results of a Probit model explaining the decision to select into the claw-back regime for Task 2. Columns (i)-(iii) probe whether loss aversion, previous experience, and the nature of the task (tedious, or not) affect the choice for the claw-back, for Treatments A-C respectively.

<< Insert Table 5 about here >>

As shown by columns (i)-(iii) of Table 5, meeting the threshold in Task 1 increases the probability to opt for the claw-back regime. We obtain similar results when using a continuous measure of productivity in Task 1 rather than a dichotomous one (the number of envelopes produced; details available on request). This likely reflects confidence in own skills or productivity, and also suggests that the claw-back can be used as a screening device by the firm – it is able to attract high productivity workers if it uses a claw-back incentive design.[20] Supporting our theory, in treatments where either subjects are not experienced (Treatment A, column i) or where commitment is unimportant (Treatment C, column iii), the coefficient of *LossAversion* is negative and significant. In contrast, experienced subjects who subsequently face the 30 minutes task (Treatment B, column ii) have a *LossAversion* coefficient that is not significantly different from zero. This zero net effect suggests offsetting disutility and

[20] Additional evidence that the claw-back contract can be used as a screening device follows from the observation that the results in Table 5 are robust to excluding the Task 1 performance indicator (*Threshold Task 1 is met Y/N*); the coefficients on *Loss Aversion* remain unchanged. Hence, Table 5 shows that when controlling for the risk and loss preferences, the more productive workers are more likely to choose the claw-back than the bonus regime. In addition, our results in columns (i)-(iii) are also robust to using the number of envelopes made in Task 1 as a control variable for skill and effort rather than a binary variable indicating whether the threshold was met.

commitment value effects. As shown in columns (iv)-(vi) of Table 5, these results are robust to including additional control variables.[21] Our results are also robust to r to pooling the data with the appropriate use of dummy variables and interaction effects. Results of these additional regressions are reported in Appendix 2 (Table A2).

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 observational data so attribution rests upon additional assumptions. We regard these education results as a robustness analysis, and formulate the following two hypotheses:

Hypothesis 4a': *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.*

The 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:

[21] When we use the *kink-based* rather than the *tail-based* measure of loss aversion, we obtain qualitatively similar but statistically weaker results. Specifically, when we identify as loss averse individuals the top 25% of subjects in terms of their ratio of slope in the loss domain to slope in the gains domain, we also find that the coefficient of the loss aversion covariate varies across treatment arms in ways that are consistent with the theory. If we use the slope as a continuous variable, then it consistently has the correct sign but does not enter significantly in any of the regression models, suggesting that we are underpowered. Details of these analyses are available upon request.

Result 4a': *Inexperienced yet educated subjects are more likely to choose the claw-back regime for a heavy task.*

Support for Result 4a': 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 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 = 0.0005 according to the appropriate two-sided Equal Proportions test.[22] ■

We also probe these issues in a regression framework. In columns (iv)-(vi) of Table 5 we explored whether education matters – even when conditioning on skills and other salient characteristics of the subjects. Two things stand out. First, the earlier results are unchanged when adding controls like age, gender, tribe, and education level. Second, education predicts selection into the claw-back, but only for experienced subjects confronted with a tedious task (column v).

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 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 well-known trade-off. On the one hand, increasing the number of fixed wage values generates information about more "intermediate points" on the aggregate demand for the claw-back, enabling more precise statements about welfare. Alternatively, this approach comes at the expense of statistical power. To test whether the demand function is linear or non-linear (either convex or concave), the literature 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 follow this approach since it is also in

[22] Interestingly, education and experience may be complements in the sophistication process. Performing the test of Result 4a on experienced subjects from Treatment B, we find that 46% of the higher educated subjects self-selected into the claw-back regime for Task 2, as opposed to 36% of the lower-educated subjects. This difference is not statistically significant (*p*-value is 0.14).

line with the spirit of our theory. 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 thus 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 97% (for a wage of UGS 150) via 51% (UGS 1200) to 21% (UGS 2400). We find that at a fixed wage of 1200 shillings the share of subjects preferring the claw-back is 51%, which means that according to our estimations the demand function for the claw-back is nearly linear. The median value is UGS 1200 and – assuming a linear demand curve – the average is UGS 1445.[23] 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 fit through three fixed wages, we present the following result:

Result 5: *The overall welfare costs incurred by experienced subjects of being offered a take-it-or-leave-it claw-back contract appear very small.*

Support for Result 5: 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 (UGS 1445). The net welfare cost associated with (forced) participation in a claw-back regime is therefore negligible for a sample of experienced workers. The small magnitude of the welfare cost suggests that for the average sophisticated worker the potential loss due to loss aversion is (nearly) offset by the commitment value of the claw-back. ■

[23] These numbers are based on a "demand curve" estimated by regressing the percentage of subjects accepting the fixed wage on different fixed wages (for the full sample of 400 subjects in Treatment D).

The finding that the average worker is nearly indifferent between the claw-back contract and a fixed-wage contract paying its expected value suggests that the marginal cost of supplying effort must be low. This follows from the observation that expected earnings are similar but subjects supply twice as much effort in the claw-back regime (on average 412.5 grams) as in the fixed wage regime (on average 187.3 grams – see Appendix 2, 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.

What are the welfare effects of introducing a *non-voluntary* claw-back contract for a sample of workers? While Result 5 suggests the average welfare effect is negligible, this zero aggregate effect may hide considerable heterogeneity. Loss neutral workers are unaffected, and neither suffer a reduction in utility when payoffs fall short of expectations, nor benefit from the claw-back's commitment value. Assume the subpopulation of loss averse workers consists of sophisticated and non-sophisticated individuals. As demonstrated by Kaur et al. (2015) and Beshears et al. (2015), some sophisticated subjects will try to leverage the claw-back's commitment value and choose a "dominated contract" (as perceived through a non-behavioral lens). These workers are better off in expected value terms. The short-term welfare implications for naïve workers are less benign, especially if there is variation in productivity across subjects, as in our sample. When we consider the subsample of subjects assigned to choose between the very low fixed wage of 150 and the claw-back, almost everybody chose the claw-back (97%). Their performance is our best approximation for the outcomes that would eventuate with a non-voluntary claw-back. For this near-universe of subjects, more than 40% did not make the threshold and therefore suffered a utility loss.

Another important consideration warrants discussion. We find that subjects are able to predict reasonably their future productivity and performance. 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 73% when the top half chooses the claw-back, and to more than 80% when the most confident 21% of subjects choose the claw-back. This pattern presents an interesting opportunity: a voluntary claw-back may be used by employers as a screening device to select the most productive workers. We hope that future work expands and focuses on this insight.

Finally, it is possible to consider overall welfare – aggregating effects across "firms" and workers. Consider the productivity effect first. Table 6 shows the Task 2 productivity effects of subjects self-selecting into the claw-back regime, for Treatments A-B. For the firm, the average number of grams sorted under the claw-back regime is 400.93 grams, and under the bonus regime it is 337.98 grams. This difference is significant at $p = 0.0001$, according to a two-sided *t*-test. The share of subjects meeting the threshold to obtain or keep the UGS 2500 reward is 0.756 for those who choose the bonus regime, and 0.914 for those who self-selected into the claw-back regime. This difference is significant at $p = 0.0045$, 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 0.175 for the claw-back regime, compared to UGS 0.178 with endogenous selection into the bonus regime.

While costs per unit of output are lower for the employer using the claw-back, we were not aiming to develop a payment regime that maximizes profits for the employer. Instead, we focused on how workers are affected by the claw-back. Indeed, since producer surplus does not monotonically increase in worker effort (the quantity of envelopes produced, or beans sorted), under our approach it was even possible for employers to earn *less* when workers put in more effort. While that did not occur, the claw-back increased productivity but only slightly improved profits. Of course, using payoff structures that strongly benefit the firm, as in Hossain and List (2012), firm profits will be enhanced with increased productivity. In this case, the claw-back regime would generate positive aggregate welfare effects.

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 behavioral expectations. More generally, the welfare implications of nudges and commitment failures remain ill-understood. As governments around the world increasingly use behavioral 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 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 – more productive workers are attracted to it, and are more likely to self-select into the loss regime when given the choice between different incentive regimes. Future work should explore the screening effect 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 behavioral nudges can be used to motivate agents without being unraveled by the dismal side of the incentive. More work is necessary, as our study represents one case study in the workplace. In this spirit, we hope that our study can represent a playbook for future advances in this area.

References

Abeler, J., A. Falk, L. Goette and D. Huffman, 2011. Reference Points and Effort Provision. *American Economic Review* 101(2): 470-492.

Allcott, H. and J. Kessler, 2015. The Welfare Effects of Nudges: A Case Study of Energy Use Social Comparisons, working paper.

Ashraf, N., D. Karlan and W. Yin, 2006. Tying Odysseus to the Mast: Evidence from a Commitment Savings Product in the Philippines. *The Quarterly Journal of Economics* 121(2): 635-672.

Augenblick, N., M. Niederle and C. Sprenger, 2015. Working over Time: Dynamic Inconsistency in Real Effort Tasks. *The Quarterly Journal of Economics* 130(3): 1067-1115

Benjamin, D.J., S.A. Brown, and J.M. Shapiro, 2013. Who is "Behavioral"? Cognitive Ability and Anomalous Preferences. *Journal of the European Economic Association* 11: 1231–1255.

Beshears, J., J. Choi, C. Harris, D. Laibson, B. Madrian and J. Sakong, 2015. Self-Control and Commitment: Can Decreasing the Liquidity of a Savings Account Increase Deposits? NBER Working paper 21474, Cambridge, MA.

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

Bryan, G., D. Karlan and S. Nelson, 2010. Commitment Devices. *Annual Review of Economics* 2: 671-698.

Camerer, C., L. Babcock, G. Loewenstein and R. Thaler, 1997. Labor Supply of New York City Cabdrivers: One Day at a Time. *Quarterly Journal of Economics* 112(2): 407-441.

Carroll, G.D., J. Choi, D. Laibson, B. Madrian and A. Metrick, 2009. Optimal Defaults and Active Decisions. *Quarterly Journal of Economics* 124(4): 1639-1674.

Chetty, R. A. Looney and K. Kroft, 2009. Salience and Taxation: Theory and Evidence. *American Economic Review* 99(4): 1145-1177.

Cohn, A., E. Fehr and L. Goette, 2015. Fair wages and effort provision: Combining Evidence from the Lab and the Field. *Management Science*, in press.

Crawford, V. and J. Meng, 2011. New York City Cabdrivers' Labor Supply Revisited: Reference-Dependence Preferences with Rational Expectations Targets for Hours and Income. *American Economic Review* 101(5): 1912-1932.

DellaVigna, S., 2009. Psychology and Economics: Evidence from the Field. *Journal of Economic Literature* 47(2): 315-372.

DellaVigna, S. and U. Malmendier, 2004. Contract Design and Self-Control: Theory and Evidence. *Quarterly Journal of Economics* 119(2): 353-402.

DellaVigna, S. and U. Malmendier, 2006. Paying Not to Go to the Gym. *American Economic Review* 96(3): 694-719.

DellaVigna, S., J. List and U. Malmendier, 2012. Testing for Altruism and Social Pressure in Charitable Giving. *Quarterly Journal of Economics* 127(1): 1-56.

DellaVigna, S. and D. Pope, 2016. What Motivates Effort? Evidence and Expert Forecasts. NBER Working Paper 22193. Cambridge, MA.

De Meza, D. and D. Webb, 2007. Incentive Design under Loss Aversion. *Journal of the European Economic Association* 5(1): 285-318.

De Quidt, J., 2017. Your Loss is my Gain: A Recruitment Experiment with Framed Incentives. *Journal of the European Economic Association* 16(2): 522-559.

Engelmann, D. and G. Hollard, 2010. Reconsidering the Effect of Market Experience on the "Endowment Effect". *Econometrica* 78(6): 2005-2019.

Fehr, E. and L. Goette, 2007. Do Workers Work More when Wages are High? Evidence from a Randomized Field Experiment. *American Economic Review* 97(1): 298-317.

Frederick, S., G. Loewenstein, and T. O'Donoghue, 2002. Time Discounting and Time Preference: A Critical Review. *Journal of Economic Literature* 40(2): 351- 401.

Fryer, R., S. Levitt, J. List and S. Sadoff, 2012. Enhancing the Efficacy of Teacher Incentives Through Loss Aversion: A Field Experiment. NBER Working Paper 18237.

Fudenberg, D. and D. Levine, 2006. A Dual-Self Model of Impulse Control. *American Economic Review* 96(5): 1449-1476.

Genesove, D. and C. Mayer, 2001. Loss Aversion and Seller Behavior: Evidence from the Housing Market. *Quarterly Journal of Economics* 116(4): 1233-1260.

Goette, L., E. Fehr and D. Huffman, 2004. Loss Aversion and Labor Supply. *Journal of the European Economic Association* 2(2-3): 216-228.

Gul, F. and W. Pesendorfer, 2001. Temptation and Self-Control. *Econometrica* 69(6): 1403-1435.

Gul, F. and W. Pesendorfer, 2004. Self-Control and the Theory of Consumption. *Econometrica* 72(1): 119-158.

Hallsworth, M., J.A. List, R. Metcalfe and I. Vlaev, 2014. The Behavioralist as Tax Collector: Using Natural Field Experiments to Enhance Tax Compliance. NBER Working Paper 2007.

Hanemann, M., 1991. Willingness to Pay and Willingness to Accept: How Much Can they Differ? *American Economic Review* 81(3): 635-647.

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

Harbaugh, W.T., K. Krause, and L. Vesterlund, 2002. Risk Attitudes of Children and Adults: Choices over Small and Large Probability Gains and Losses. *Experimental Economics* 5(1): 53–84.

Henrich, J., S.J. Heine and A. Norenzayan, 2010. The weirdest people in the world? *Behavioral and Brain Sciences* 33: 61–135.

Herweg, F. and K. Mierendorff, 2013. Uncertain Demand, Consumer Loss Aversion and Flat-Rate Tariffs. *Journal of the European Economic Association* 11(2): 399-432.

Herweg, F., D. Muller and P. Weinschenk, 2010. Binary Payment Regimes: Moral Hazard and Loss Aversion. *American Economic Review* 100(5): 2451-2477.

Hossain, T. and J.A. List, 2012. The Behavioralist Visits the Factory: Increasing Productivity Using Simple Framing Manipulations. *Management Science* 58(12): 2151-2167.

Imas, A., S. Sadoff, and A. Samek, 2016. Do People Anticipate Loss aversion? *Management Science* 63(5): 1271-1284.

Kahnemann, D., J. Knetsch and R. Thaler, 1990. Experimental Tests of the Endowment Effect and the Coase Theorem. *Journal of Political Economy* 98(6): 1325-1348.

Kaur, S., M. Kremer and S, Mullainathan, 2015. Self-Control at Work. *Journal of Political Economy* 123(6): 1227-1277

Köszegi, B. 2014. Behavioral Contract Theory. *Journal of Economic Literature* 52(4): 1075-1118.

Laibson,, D., 1997. Golden Eggs and Hyperbolic Discounting. *Quarterly Journal of Economics* 112: 443-477.

Levitt, S., J.A. List, S. Neckermann and S. Sadoff, 2012. The Behavioralist Goes to School: Leveraging Behavioral Economics to Improve Educational Performance, NBER Working Paper 18165.

List, J.A., 2013. Does Market Experience Eliminate Market Anomalies? *Quarterly Journal of Economics* 118(1): 41-71.

List, J.A., 2004. Neoclassical Theory Versus Prospect Theory: Evidence from the Marketplace. *Econometrica* 72(2): 615-625.

List, J.A., 2011. Does Market Experience Eliminate Market Anomalies? The Case of Exogenous Market Experience. *American Economic Review P&P* 101 (3): 313-317.

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

List, J.A. and A. Samek, 2015. The Behavioralist as Dietician: Leveraging Behavioral Economics to Improve Child Food Choice and Consumption. *Journal of Health Economics* 39: 135-146.

Luft, J., 1994. Bonus and Penalty Incentives Contract Choice by Employees. *Journal of Accounting and Economics* 18(2): 181-206

Mani, A., S. Mullainathan, E. Shafir, and J. Zhao, 2013. Poverty Impedes Cognitive Function. *Science* 341 (6149): 976-980.

McClure, S., D. Laibson, G. Loewenstein and J. Cohen, 2004. Separate Neural Systems Value Immediate and Delayed Monetary Rewards. *Science* 306: 503-507.

O'Donoghue, T. and M. Rabin, 1999. Doing It Now or Later. *American Economic Review* 89(1): 103-124.

Rabin, M., 1998. Psychology and Economics. *Journal of Economic Literature* 36(1): 11-46.

Samuelson, W. and R. Zeckhauser, 1988. Status Quo Bias in Decision-Making. *Journal of Risk and Uncertainty* 1(1): 7-59.

Shiv, B. and A. Fedorikhin, 1999. Heart and Mind in Conflict: The Interplay of Affect and Cognition in Consumer Decision-Making. *Journal of Consumer Research* 26(3): 278-292.

Thaler, R., 1980. Toward a Positive Theory of Consumer Choice. *Journal of Economic Behavior & Organization* 1(1): 39-60.

Thaler, R. and H. Shefrin, 1981. An Economic Theory of Self Control. *Journal of Political Economy* 89(2): 392-406.

Thaler, R. and C.R. Sunstein, 2008. Nudge: Improving Decisions about Health, Wealth, and Happiness, Yale University Press.

Trope, Y. and A. Fishbach, 2000. Counteractive Self-Control in Overcoming Temptation. *Journal of Personality and Social Psychology* 79(4): 493-506.

APPENDIX 1: A MOTIVATING MODEL

In this Appendix we formulate and solve a model based on an intra-individual principal-agent problem. This model yields most of the hypotheses we set out to test in the main text, but of course it is possible to write up alternative models producing similar conclusions.[24] The lack of a one-on-one matching to theory implies the empirical section cannot "prove" the theory. Instead, we regard the theory as a lens through which we might approach our data – it is a motivational device. We use backward induction to solve this model, and start with the biased agent's problem.

A.1 The agent's effort decision

Assume an agent has the following utility function, which displays both the self-control bias and loss aversion or reference-dependent utility:

$$u = \begin{cases} w - \alpha c(e), & if \quad w \geq r \\ w + \theta(w - r) - \alpha c(e), & if \quad w < r \end{cases} \tag{1}$$

where $w \geq 0$ are the actual earnings, $r \geq 0$ is the reference value used by the agent to compare his earnings to, e equals effort in a production task (see below), $c(e)$ are convex effort costs, and (θ, α) are parameters. Loss aversion implies earnings (outcomes) below a reference value induce a loss in utility ($\theta \geq 0$), and we assume the magnitude of the loss is linear in the distance between the realized outcome and the reference value. Similar to O'Donoghue and Rabin (1999) the agent's lack of self-control is captured by $\alpha \geq 1$, 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.

In case $\theta \geq 0$, the specification in (1) 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. Normalizing the principal's relative evaluation of earnings and effort costs to 1, the individual has a self-control problem if $\alpha > 1$ – the "planning me" (or principal) values the future benefits of receiving (or keeping) the reward more highly than the "implementing me" (or agent). This is a familiar problem studied for

[24] 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)).

example in the context of a health-conscious principal that seeks to motivate the shirking agent to go to the gym (e.g. DellaVigna and Malmendier (2006)).[25]

Assume that the agent is invited to engage in a task with two possible outcomes. If production equals or exceeds an exogenous (and unknown) threshold T, the agent receives a payment equal to $R > 0$. In contrast, when production is below the threshold, the agent receives nothing, so that $w = \{0,R\}$. Production is a positive function of effort, e, and the probability of meeting threshold T is denoted by $p(e)$, with $p'(e) > 0$. If the agent's reference value for a certain task equals zero ($r = 0$) in equation (1), then the agent maximizes the following function:

$$MAX_e\, u = p(e)R - \alpha c(e), \qquad (2)$$

and the agent's optimal effort level, e^B, implicitly solves

$$c'(e) = p'(e)R/\alpha. \qquad (2')$$

Note that e^B is a not function of θ (as payoffs are by definition in the gains domain as $r = 0$ in this case), but that it is a function of α. Assuming that $p'' \leq 0$ and $c'' > 0$, the more the agent is inclined to over-estimate the cost associated with providing effort, the lower the optimal effort level as chosen by the agent – the self-control problem ($de^B/d\alpha < 0$).

There are cases where the agent might have a different reference value. For example, and for reasons explained more fully below, the agent's reference value might equal the potential payment ($r = R$). If so, using (1) the agent maximizes:

$$MAX_e\, u = p(e)R - \big(1 - p(e)\big)R\theta - \alpha c(e), \qquad (3)$$

and the agent's optimal effort level, e^C, implicitly solves

$$c'(e) = p'(e)R(1 + \theta)/\alpha. \qquad (3')$$

Here, e^C is a function of both α and θ. Comparing (2') and (3') it is immediately clear that for any $\alpha \geq 1$ we have $e^C(\alpha, \theta) > e^B(\alpha)$ if $\theta > 0$ – loss averse agents will "work harder" to avoid the loss associated with giving up the payment, and the difference in effort levels is larger the higher is θ.

[25] Note that, in principle, the individual also faces a problem if $\alpha < 1$. In that case the agent over-invests in effort compared to the principal's preferences.

Imagine a so-called "*bonus regime*" where the agent receives payment R in case she meets a certain performance threshold. We assume such ex post payments do not instill great sentiments of ex-ante ownership, hence $r = 0$ and the agent chooses $e^B(\alpha)$. In contrast, the principal may also select a so-called "*claw-back regime.*" This regime is described by ex-ante payments equal to R, but now the agent has to return the payment in case productivity falls below the threshold. Prior experiments (see List (2003), (2004)) have revealed that ex ante transfers create sentiments of ownership, so $r = R$ and the agent chooses $e^C(\alpha, \theta)$.[26]

A.2 The principal's welfare levels under the two regimes

We first derive the consequences of the agent's effort decisions for the principal's welfare. As stated above, we assume the principal does not suffer from any behavioral biases. Using (1) and setting $\theta = 0$ and $\alpha = 1$, the principal's welfare level[27] is

$$V = p(e)R - c(e)). \tag{4}$$

The agent's effort level that maximizes the principal's welfare, e^*, implicitly solves

$$c'(e) = p'(e)R. \tag{4'}$$

We use $V^* = V(e^*)$ to denote the maximum welfare the principal can attain, and $V^B = V\left(e^B(\alpha)\right)$ and $V^C = V\left(e^C(\alpha, \theta)\right)$ as the level of welfare 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 A.1: *For any $\theta > 0$, there is a critical level of α, $\bar{\alpha}(\theta)$, where the principal is indifferent between the bonus and claw-back regime. The principal's welfare is higher in the bonus regime than in the claw-back regime for $\alpha < \bar{\alpha}(\theta)$, and the opposite holds if $\alpha > \bar{\alpha}(\theta)$.*

[26] In studies of the claw-back that involve a significant delay in repaying the money (e.g. as in Fryer et al. (2012)) 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, and in this time period the moneys cannot be spent anyway.

[27] It is common to assume that the principal maximizes her own payoff 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.

Proof. From first-order conditions (2')-(4') it immediately follows that $V^B(\alpha) = V^*$ if $\alpha = 1$, and $V^C(\alpha, \theta) = V^*$ if $\alpha = 1 + \theta$. Next, $\frac{dV^B}{d\alpha} = [p'R - c']\frac{de^B}{d\alpha} = \frac{(p'R-c')c'}{p"R-\alpha c"}$. Here, $c' > 0$ and the denominator is negative because of the second-order condition of profit maximization. Comparing (2') and (4') we have $e^B(\alpha) < e^*$ for all $\alpha > 1$, and hence $\frac{dV^B}{d\alpha} < 0$ for all $\alpha > 1$. Similarly, we have $\frac{dV^C}{d\alpha} = [p'R - c']\frac{de^C}{d\alpha} = \frac{(p'R-c')c'}{(1+\theta)p"R-\alpha c"}$. Comparing (3') and (4') we have $e^C(\alpha, \theta) > e^*$ if $\alpha < 1 + \theta$, and $e^C(\alpha, \theta) < e^*$ if $\alpha > 1 + \theta$. Hence 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 $1 < \alpha$ and (iii) $\frac{dV^C}{d\alpha} > 0$ for $1 \leq \alpha < 1 + \theta$, there exists a critical level of α, $1 \leq \bar{\alpha}(\theta) < 1 + \theta$, such that $V^C(\alpha, \theta) < V^B(\alpha)$ for $\alpha < \bar{\alpha}(\theta)$ and $V^C(\alpha, \theta) > V^B(\alpha)$ for $\alpha > \bar{\alpha}(\theta)$. ∎

Lemma A.1 is illustrated in Figure A.1, which plots the principal's welfare for a range of (lack of) self-control values of the agent (α), if the agent works under a bonus regime or under a claw-back regime. The welfare levels under the claw-back regime are depicted for two levels of loss aversion (θ^L, θ^H, with $\theta^H > \theta^L$). In the bonus regime, the agent puts in a level of effort equal to the effort that maximizes the principal's welfare if α = 1 (and hence $V^B(\alpha) = V^*$ if $\alpha = 1$). He will put in less effort when α > 1. The difference between e^* and $e^B(\alpha)$ increases with α, and the principal's welfare level $V^B(\alpha)$ is a monotonically decreasing function of α.

If a loss-averse agent without a self-control problem (i.e. $\alpha = 1$) 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 α > 1, the difference between e^* and $e^C(\alpha, \theta^j)$, $j = \{L, H\}$, first decreases, becomes zero (at $\alpha = 1 + \theta^j$), and then becomes more and more negative. This means that $V^C(\alpha, \theta^j)$ is a hump-shaped function of α, resulting in $V^C(\alpha, \theta^j) = V^*$ at $\alpha = 1 + \theta^j$. Combining, for given θ the principal prefers to select his agent into the claw-back regime (as opposed to the bonus regime) if $\alpha > \bar{\alpha}(\theta)$. As

shown in Figure A.1 we have $d\bar{\alpha}(\theta)/d\theta > 0$, and the optimal choice for the principal depends on the combination of α and θ.[28]

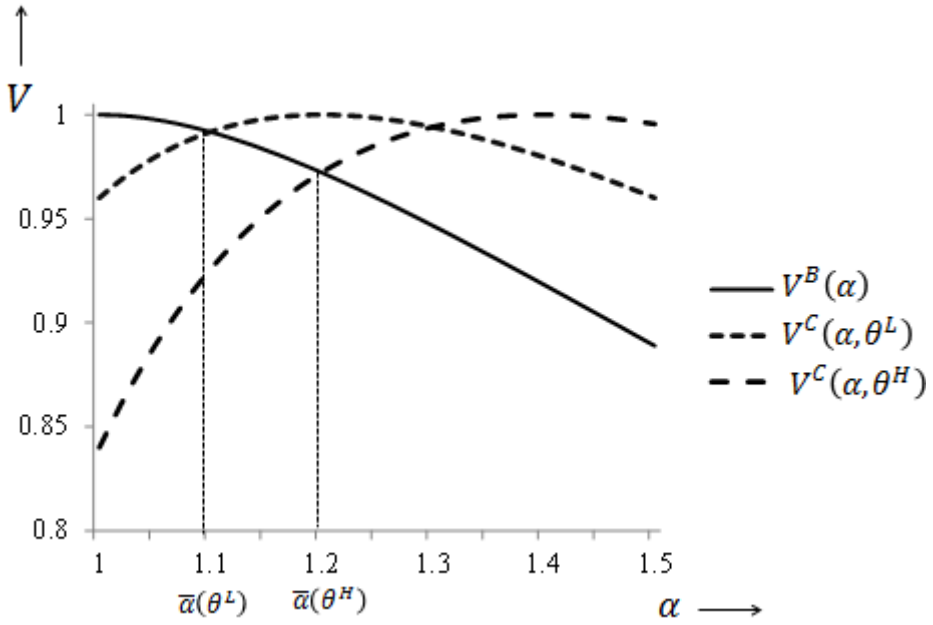

Figure A1: Payoffs for the principals as a function of α under the two incentive regimes, and for two different degrees of loss aversion for the agent.

A.3 Experience, education and sophistication

In light of the pay-offs for the Principal as depicted in Figure A1, which payment regime should the principal choose? All principals know the levels of α and θ of "their agent." Since naïve principals do not realize that loss aversion can be leveraged to overcome the self-control problem, such principals assume that behavior of their agent is governed by (2'). While this principal's welfare is identical under the claw-back and the bonus we assume he weakly prefers the bonus (to avoid disutility to the agent). A sophisticated principal, instead, will choose the

[28] Figure A.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 $\frac{\partial\bar{\alpha}(\theta)}{\partial\theta} > 0$, $\bar{\alpha}(\theta^L) = 1.10$ and $\bar{\alpha}(\theta^H) = 1.20$. Alternatively, we can write $\bar{\theta}(\alpha) = 2(\alpha - 1)$, and then the principal prefers the claw-back over the bonus regime if $\theta < \theta(\alpha)$. If $\theta > \theta(\alpha)$, the agent works too hard under the claw-back, and the principal prefers to select him into the bonus regime.

bonus or claw-back, depending on which regime delivers the greatest welfare level (i.e. depending on θ and α, as illustrated in Figure A1). Principals are either naïve or sophisticated, so for any principal the following holds:

$$V = I(S)[\max(V^B, V^C), 0] + (1-I(S))V^B \quad (5)$$

where I is an indicator function taking the values of 0 or 1, indicating whether the principal is sophisticated or not. We assume sophistication, S, is a latent variable that depends on experience with the claw-back, E, and possibly other variables such as education, Z. For example, one convenient representation, where subscript i indexes subjects, is as follows:

$$S_i^* = \varphi_0 + \varphi_1 E_i + \varphi_2 Z_i + \varepsilon_i. \quad (6)$$

The idea is that there exists a continuous but unobservable function that captures the worker's ability to "think through" the implications of payment regimes, and that this function takes on higher values because of experience or education.[29] Somebody is sophisticated if:

$$S_i = \begin{cases} 1 & if\ S_i^* > 0 \\ 0 & if\ S_i^* \leq 0 \end{cases}.$$

We assume I_i=1 for S_i=1 and I_i=0 for S_i=0.

[29] Note that education is expected to increase one's cognitive skills, but of course it may also act as a "filter" – only those workers that have sufficient self-control, are able to complete their education. This would imply that educated workers are *less* likely to select the claw-back regime, as they have less need for a (soft) commitment device. However, as shown in Table 5 we find that educated subjects are more likely to choose the claw-back regime (conditional on previous exposure to the claw-back), and hence the "cognitive skills" effect dominates the effect of "being less prone to self-control issues".

APPENDIX 2

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

Table A2: Probit regression results based on pooled data Treatment arms A, B and C

	(1) Choose claw-back regime	(2) Choose claw-back regime	(3) Choose claw-back regime
Treatment B	0.370**	0.424***	0.594***
	(0.145)	(0.143)	(0.169)
Treatment C	0.0279	0.130	0.204
	(0.150)	(0.142)	(0.161)
#Envelopes Folded in Task (1)	0.0238***		
	(0.00702)		
Threshold Task 1 met (y/n)		0.543***	0.544***
		(0.148)	(0.149)
LossAversion	-0.244***	-0.267***	-0.514***
	(0.0851)	(0.0851)	(0.174)
LossAversion × Treatment B			0.408*
			(0.215)
LossAversion × Treatment C			0.245
			(0.219)
Risk Aversion	-0.116*	-0.116*	-0.121**
	(0.0596)	(0.0594)	(0.0590)
Tribe other than Musoga	0.359+	0.386*	0.433*
	(0.231)	(0.229)	(0.231)
Education level	0.0639	0.0675	0.0707*
	(0.0414)	(0.0416)	(0.0420)
Age	-0.00521	-0.00538	-0.00547
	(0.00402)	(0.00403)	(0.00401)
Female	-0.0358	-0.0105	-0.0216
	(0.113)	(0.112)	(0.115)
_cons	-1.590***	-1.536***	-1.655***
	(0.345)	(0.328)	(0.344)
N	600	600	600

Standard errors in parentheses, * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

TABLES

Table 1: Experimental design

<table>
<tr><th>Treatment A</th><th colspan="2">Treatment B</th><th colspan="2">Treatment C</th><th colspan="2">Treatment D</th></tr>
<tr><td colspan="7">Task 1, producing envelopes</td></tr>
<tr><td>bonus regime

N=200</td><td colspan="2">claw-back regime

N=200</td><td colspan="2">claw-back regime

N=200</td><td colspan="2">claw-back regime

N=400</td></tr>
<tr><td>30 min envelope folding (A1)

Receive payment if $Q_1 \geq T_1$</td><td colspan="2">30 min envelope folding (B1)

Return payment if $Q_1 < T_1$</td><td colspan="2">30 min envelope folding (C1)

Return payment if $Q_1 < T_1$</td><td colspan="2">30 min envelope folding (D1)

Return payment if $Q_1 < T_1$</td></tr>
<tr><td colspan="7">Task 2, sorting beans</td></tr>
<tr><td colspan="2">Self-selection in one of following regimes</td><td colspan="2">Self-selection in one of following regimes</td><td colspan="2">Self-selection in one of following regimes</td><td colspan="2">Self-selection in one of following regimes</td></tr>
<tr><td>bonus regime N=?</td><td>claw-back regime N=?</td><td>bonus regime N=?</td><td>claw-back regime N=?</td><td>bonus regime N=?</td><td>claw-back regime N=?</td><td>Fixed wage N=?</td><td>claw-back regime N=?</td></tr>
<tr><td>30 min bean sorting (A21)

Receive bonus if $Q_2 \geq T_2$</td><td>30 min bean sorting (A22)

Return bonus if $Q_2 < T_2$</td><td>30 min bean sorting (B21)

Receive bonus if $Q_2 \geq T_2$</td><td>30 min bean sorting (B22)

Return bonus if $Q_2 < T_2$</td><td>3 min bean sorting (C21)

Receive bonus if $Q_2 \geq T_3$</td><td>3 min bean sorting (C22)

Return bonus if $Q_2 < T_3$</td><td>30 min bean sorting (D21)

Receive Y∈{150, 1200,2400}</td><td>30 min bean sorting (D22)

Return bonus if $Q_2 < T_2$</td></tr>
</table>

Table 2: Summary statistics.

Variable	N	mean	median	sd
DegreeLossAversion	1200	-0.38038	0	0.730156
DegreeRiskAversion	1200	0.186292	0	0.976409
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 = 0.0080
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)	0.536**	0.743***	0.481*	0.549**	0.696**	0.435*
	(0.241)	(0.258)	(0.256)	(0.251)	(0.283)	(0.266)
RiskAversion	-0.254**	0.00579	-0.181*	-0.265**	0.00702	-0.191*
	(0.127)	(0.0918)	(0.103)	(0.128)	(0.0914)	(0.103)
LossAversion	-0.541***	-0.0972	-0.326**	-0.582***	0.00425	-0.291**
	(0.181)	(0.131)	(0.137)	(0.186)	(0.139)	(0.141)
Tribe other than Musoga				0.339	0.680*	0.0420
				(0.377)	(0.361)	(0.545)
Education level				0.0367	0.159**	0.0589
				(0.0648)	(0.0787)	(0.0789)
Age				-0.0003	-0.0122**	-0.0003
				(0.00833)	(0.00591)	(0.00750)
Female				0.103	-0.335*	0.186
				(0.217)	(0.203)	(0.209)
Constant	-1.280***	-0.924***	-1.057***	-1.751***	-1.235**	-1.275*
	(0.222)	(0.242)	(0.237)	(0.567)	(0.574)	(0.681)
N	200	200	200	200	200	200
Wald Chi2	14.93***	9.42**	10.68**	16.19**	23.72***	12.09*
Pseudo-R2	0.08	0.04	0.04	0.09	0.10	0.05

Table 6: Grams of beans sorted in Task 2 of treatments A and B, for different regimes.[a]

		Regime chosen for Task 2 (sorting beans)		
		bonus	claw-back	
	Treatment A	368.72 (108.18) n = 154	371.61 (93.28) n = 46	p = 0.870
	Treatment B	337.98 (119.16) n = 119	400.93 (86.78) n = 81	p < 0.000

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