UNIVERSITÉ DE FRIBOURG
UNIVERSITÄT FREIBURG

Faculty of Economics and Social Sciences
Department of Economics
Chair of Industrial Economics

University of Fribourg

05 Pre-analysis plan

22.10.2018

Hidden costs of control: evidence from the field

Author:

Prof. Dr. Holger Herz

holger.herz@unifr.ch

Author:

Christian Zihlmann

christian.zihlmann@unifr.ch

Contents

1 Introduction

1.1 Abstract

The purpose of this research project is to find field evidence of hidden costs of control as they were found in the lab by Falk and Kosfeld (2006) (henceforth: F&K). Moreover, we aim to uncover heterogeneity in the population through elicitation of agents' social preferences in a subsequent stage; and match social preferences to the behavioral response observed in the field.

1.2 Motivation

In a principal-agent setting, F&K analyze the consequences of control on agent's motivation. Contrary to agency theory, they find that most agents reduce their effort as a response to the principal's decision to implement a minimum performance requirement. As a consequence, control entails hidden costs. The existence of hidden costs of control has been replicated in the lab in many variations and instances[1]. However, field evidence is very rare and does not mimic F&K's lab experiment precisely. Nagin, Rebitzer, Sanders, and Taylor (2002) conduct a field experiment in a call-center company. They analyze the impact of monitoring on workers behavior, where the controlling device is a monitoring rate defined as the frequency workers' calls were tested for correctness. Boly (2011) conducts a field experiment too, where the controlling device is an audit of the work output, occurring with a certain probability. In case of non-compliance, workers are fined and receive a penalty. Both studies find standard agency theory to be supported. Belot and Schröder (2016) conduct a framed field experiment with students where subjects are aware that there is a study going on[2]. Also this study implements an audit mechanism as monitoring device, accompanied by penalties in case of non-compliance. The study documents negative spillover effects to another productivity dimension than the contracted dimension, supposing that agents choose the cheapest way for reciprocating their negative behavioral reaction. As seen, all three field studies implement an audit as controlling device, which is in contrast to F&K who implement a minimum performance requirement. Audits may add complexity to individual decision-making due to probabilistic reasoning. Also, these three studies are only able to control for individual heterogeneity within clear limits. Lastly, the findings do not reconcile.

This project shall address these deficiencies. The proposed field experiment contributes to

[1] See Falk and Kosfeld (2006); Dickinson and Villeval (2008); Schnedler and Vadovic (2011); Ziegelmeyer, Schmelz, and Ploner (2012); Kessler and Leider (2013); Masella, Meier, and Zahn (2014); Schmelz and Ziegelmeyer (2015); Kessler and Leider (2016); Riener and Wiederhold (2016); Burdin, Halliday, and Landini (2018).

[2] While the task is novel and clever, it is also relatively artificial: identifying the value and country of euro coins.

the literature in several ways: First of all, we mirror F&K's experimental design as closely as possible and implement a minimum performance requirement as controlling device. Second, our experiment is conducted in a real labor market and hence qualifies as a natural field experiment according to Harrison and List (2004). Third, we employ a large sample size from a real labor market, being more representative than the student populations used in previous studies (Boly, 2011; Belot & Schröder, 2016). Lastly and importantly, we elicit social preferences and are thus able to identify heterogeneous treatment effects, possibly allowing us to drive conclusions about the behavioral driver of such a negative reaction.

The main purpose of our experiment is to evaluate the external validity of F&K's findings by investigating if hidden costs of control exist in the field, assessing the magnitude, and analyzing heterogeneous treatment effects with regard to social preferences.

1.3 Research Questions

1. Do hidden costs of control exist in the field and if so, do hidden costs of control outweigh the direct benefit of control?

2. Heterogeneity: are hidden costs of control larger when agents are intrinsically motivated?

3. Heterogeneity: are hidden costs of control positively correlated with a preference for negative reciprocity?

4. Heterogeneity: are hidden costs of control positively correlated with a preference for positive reciprocity?

5. Heterogeneity: are hidden costs of control positively correlated with a preference for trust?

2 Research Strategy

2.1 Experimental Design

This study is divided into three separate tasks classified into two parts: Part A includes the field experiment with two separate real-effort tasks to elicit workers effort, while the subsequent Part B elicits workers social preferences[3]. Figure 1 summarizes the experimental design.

[3] All parts of the experiment are coded with the software Otree (Chen, Schonger, & Wickens, 2016). Otree has an integrated interface for AMT. Instructions will be published and code will be available on request.

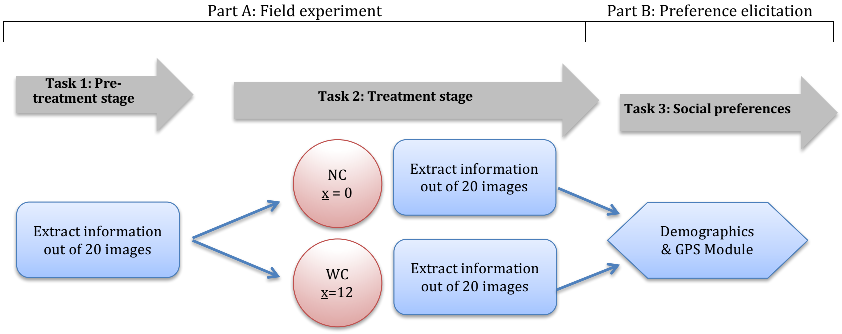
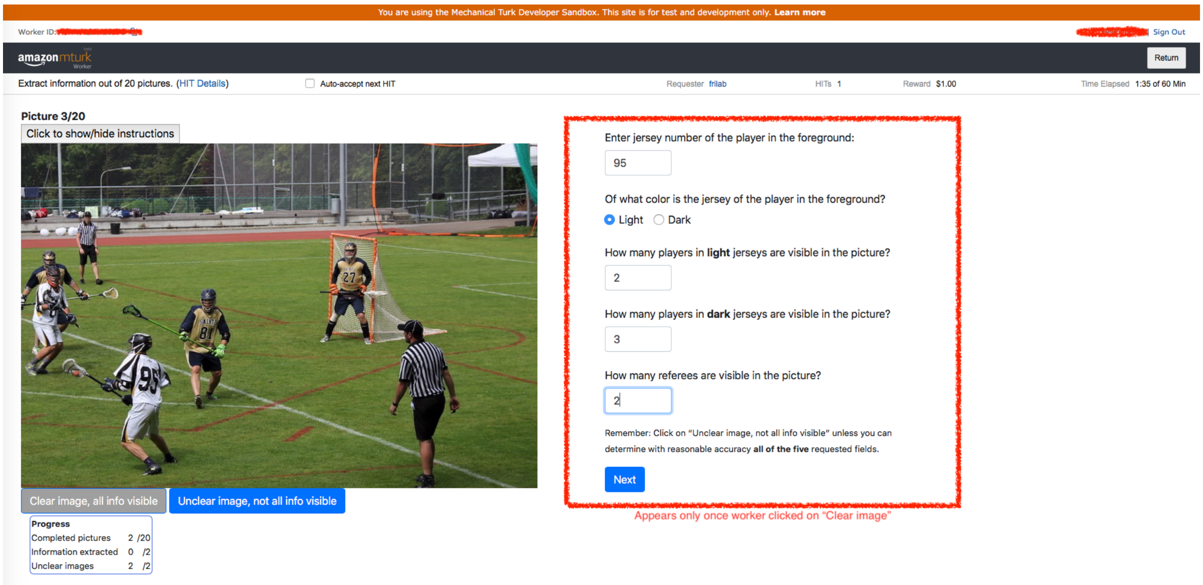

Part A: Field experiment
Part B: Preference elicitation
Task 1: Pre-treatment stage
Task 2: Treatment stage
Task 3: Social preferences
Extract information out of 20 images
NC
x = 0
WC
x=12
Extract information out of 20 images
Extract information out of 20 images
Demographics & GPS Module

Figure 1: Experimental design illustrated in a flowchart

Part A: Field experiment

The experiment is conducted in an online labor market intermediary: we recruit workers through Amazon Mechanical Turk ("AMT"). We play the principal or employer ("requester") and offer a one-time employment contract with a fixed reward in case of a so-called Human Intelligence Task ("HIT") completion. Agents ("workers") are not aware that they participate in an experiment and engage in a naturalistic real-effort task commonly posted on AMT: extracting information out of a picture in order to categorize these. Concretely, we present workers with pictures from game-play situations of a lacrosse game. We ask workers to extract the following information out of that picture: the jersey number of the player in the foreground, the color of its jersey, the total count of light and dark colored jerseys, and the total count of referees. Pictures vary in the degree of difficulty, requiring a different degree of effort to solve. Figure 2 illustrates an easy-to-solve situation.


You are using the Mechanical Turk Developer Sandbox. This site is for test and development only. **Learn more**

Worker ID: [redacted]

[redacted] Sign Out

Extract information out of 20 pictures. (HIT Details) □ Auto-accept next HIT Requester frilab HITs 1 Reward $1.00 Time Elapsed 1:35 of 60 Min

Picture 3/20

Click to show/hide instructions

Enter jersey number of the player in the foreground:

95

Of what color is the jersey of the player in the foreground?

● Light ○ Dark

How many players in **light** jerseys are visible in the picture?

2

How many players in **dark** jerseys are visible in the picture?

3

How many referees are visible in the picture?

2

Remember: Click on "Unclear image, not all info visible" unless you can determine with reasonable accuracy **all of the five** requested fields.

Next

Appears only once worker clicked on "Clear image"

Clear image, all info visible | Unclear image, not all info visible

Progress		
Completed pictures	2	/20
Information extracted	0	/2
Unclear images	2	/2

Figure 2: Easy-to-solve picture

First, a pre-treatment stage (HIT1) is conducted where all workers are subject to a no-control environment. This stage has a two-fold purpose: first, HIT1 serves a lock-in task with the goal to reduce attrition once treatment is induced. Second, we are able to collect pre-treatment individual performance characteristics.

Workers are presented 20 pictures. For each picture, workers need to decide whether they can solve that picture. This is the case if all requested information is visible ("Clear image, all info visible"-button). Workers can also decide to opt-out. The opt-out button ("Unclear image, not all info visible"-button) is the truthful response if workers cannot solve a picture; e.g. if the picture is blurry or the requested information is not identifiable. Such an opt-out option is very commonly used on AMT and hence natural to workers. The opt-out option allows for cheap shirking since in HIT1 all workers are automatically paid regardless of their output - no worker is subject to a control mechanism. Thanks to such a button, we are able to induce variation in workers' effort, measured through the number of pictures solved. Once workers have completed all 20 images, they are paid USD 1 and are granted a qualification on AMT. With this qualification, they have the opportunity to do a different set of 20 pictures in another HIT. This is the treatment stage (HIT2) where the contract of workers is varied. The control group receives the same contract as in HIT1 and is again not subject to any control mechanism. For the treatment group, a control mechanism in the form of a minimum performance requirement $\underline{x}$ is implemented.

- NC - no MPR (control group)
  Same incomplete contract as in pre-treatment stage HIT1: no minimum performance requirement implemented ($\underline{x}$=0)[4].

- WC - low MPR
  We implement a weak, inefficient control device by setting a low minimum performance requirement allowing workers to click on the opt-out option relatively often, that is 8 times out of 20 ($\underline{x}$=12)[5].

At the end of HIT2, we elicit (i) individual fairness perceptions with regard to the reward[6] (ii) intrinsic motivation to fulfill the task by asking workers if they play or regularly watch lacrosse[7] and (iii) an additional variable controlling for the device workers are using.

[4] All work is accepted: your HIT will be approved automatically within 1 day. We do not have the possibility to review the quality of your work before approval. Nevertheless, please be as accurate and precise as possible.

[5] The count of your clicks on the "Unclear image, not all info visible"-button will be checked by the computer. Your HIT will be approved automatically when you try to solve at least 12 pictures. Namely, we will reject the HIT if you click on "Unclear image, not all info visible" more than 8 times. We do not have the possibility to review the quality of your work before approval. Hence, if you try to solve 12 pictures, your work is automatically processed for payment. Nevertheless, please be as accurate and precise as possible.

[6] inspired by Cohn, Fehr, and Goette (2015).

[7] Self-reported measures for intrinsic motivation may likely be confounded since the treatment is induced beforehand and may well affect workers intrinsic motivation, jeopardizing reliability and validity of the measure. Therefore, we argue that familiarity with lacrosse is a valid proxy for intrinsic motivation in this task: workers

Part B: Preference elicitation

Some weeks after the field experiment, we will invite all workers who completed the treatment stage to participate in an academic study (HIT3). In this stage, we will (i) collect demographic data and (ii) employ the streamlined method of the Global Preference Survey ("GPS")[8]. This stage is identical for both groups.

2.2 Sampling

Workers will be recruited from AMT. We restrict our sample to workers with a permanent residence in the U.S.. Since we want to employ a sample best representing a labor market, we do not impose further common restrictions, such as Master's qualification or a certain % of successfully completed HITs. Therefore, we expect the characteristics of our population to be relatively representative of the U.S. internet population (Ipeirotis, 2010). Workers will be randomly assigned to the treatment and control group, constituting the exogenous variation in this study. All workers who complete HIT2 will be included in the sample for conducting the statistical analysis of hypotheses 1 and 2. Hypotheses 3 to 5 require the social preference data collected in HIT3. Due to potential attrition, the according analysis will be conducted with a smaller sample, i.e. all workers who complete HIT3.

There is the theoretical possibility that workers assigned to WC do not comply with the minimum performance requirement since we cannot strictly enforce it. In this case, the HIT will be rejected as a penalty and workers do not earn the monetary reward. From a data analysis perspective, these non-compliers will be treated in two different ways to ensure robustness of the procedure and the results: first, we will simply take the observed clicks on the opt-out button (observed behavior). Second, we set their number of clicks on the opt-out button to the maximum allowed quantity, that is 8 clicks[9] to reflect a hard enforcement of the controlling device.

For the determination of sample size and power calculations as well as attrition considerations, please refer to section 4 Piloting and Power Analysis.

who play or regularly watch lacrosse are arguable more intrinsically motivated doing this task well. Furthermore, the objective binary variable if one plays lacrosse or not is reliable and not exposed to measurement error.

[8] The GPS is an experimentally validated data-set of time preference, risk preferences, positive and negative reciprocity, altruism, and trust (Falk et al., 2018; Falk, Becker, Dohmen, Huffman, & Sunde, 2016; Falk, Becker, Dohmen, Enke, & Huffman, 2015).

[9] Variable POST_OOCOM, see C Appendix: Variable definition for further details.

2.3 Main outcome variables

2.3.1 Main endogenous variable

The main dependent variable is based on the number of clicks on the "Unclear image"-button, called the opt-out option, and is a proxy for effort.

$POST\_OO_i =$ number of clicks on the opt-out option after treatment induction (task 2), being a proxy of effort respectively **shirking** for each *individual i*

For example, a $POST\_OO_i = 2$ is expected for fully honest agents (remember that we include two truly blurry pictures, where the "Unclear image"-button is the truthful response). A $POST\_OO_i > 2$ however clearly indicates shirking as all the other 18 pictures have a clear unique solution. In short, $POST\_OO_i$ is they key endogenous variable for identifying an average treatment effect.

In order to identify heterogeneous treatment effects linked with social preferences, one also needs to account for heterogeneity among the population: workers are likely heterogeneous and some may exhibit from the beginning - that is before treatment induction - a higher inclination to shirk than others. That is why we are also interested in the evolution of shirking behavior for an individual worker. The second main outcome variable ΔOO_i represents the change in the number of clicks on the opt-out button from stage 1 to stage 2.

$PRE\_OO_i =$ number of clicks on the opt-out option in the pre-treatment stage (task 1), a proxy of effort respectively **shirking** for each *individual i*

$\Delta OO_i =$ $POST\_OO_i - PRE\_OO_i$ representing the **difference in shirking frequency** between post-treatment and pre-treatment stage, for each *individual i*

For instance, a $\Delta OO_i < 0$ represents workers shirking more often in HIT1 than in HIT2[10]. A $\Delta OO_i = 0$ stands for workers who shirk equally in HIT1 and HIT2. Importantly, $\Delta OO_i > 0$ indicates an increase in shirking behavior from HIT1 to HIT2. Being a proxy for the evolution of shirking behavior from stage 1 to stage 2, ΔOO_i is the key outcome variable of interest for

[10] The task is designed that learning is very limited to occur. That is why there is no obvious reason to expect a substantial part of workers following this pattern in the no-control condition, however, in the WC condition, such a pattern would reconcile with standard agency theory: control mechanism increase workers effort by preventing them from shirking.

the analysis of heterogeneous treatment effects.

2.3.2 Main explanatory variables

The following table gives a quick overview of the main explanatory variables. A detailed summary of all variables, incl. its elicitation and computation, is enclosed in C Appendix: Variable definition on page 22.

Variable	Grouping	Description	Properties
TREATMENT	D	Indicating treatment groups	Dummy; 0 if NC and 1 if WC
LACROSSE	μ	Proxy for intrinsic motivation	Dummy; 1 if worker plays lacrosse
NEGREC	δ	Preference for negative reciprocity	standard score (GPS module)
POSREC	δ	Preference for positive reciprocity	standard score (GPS module)
TRUST	δ	Preference for trust	standard score (GPS module)

Table 1: Main explanatory variables

2.4 Hypotheses

2.4.1 Primary Hypothesis

Behavioral Prediction 1. ***Control-aversion.*** *Control-averse agents lower their work effort when the principal imposes a weak controlling device and thus exhibit higher shirking than under a no-control condition.*

Hypothesis 1. ***Existence and magnitude of hidden costs of control.*** *For the treatment group WC, shirking behavior - represented by the endogenous variable $POST\_OO_i$ - will be significantly higher compared to the control group NC.*

Let us revisit the key endogenous variable: $POST\_OO_i$ is the number of clicks on the opt-out button in the treatment stage. Workers assigned to WC experience the implementation of a control device[11]. As a consequence of Behavioral Prediction 1, we expect higher shirking in stage 2 for workers assigned to WC.

It is important to note that we clearly formulate a directional hypothesis: $POST\_OO_i$ will be significantly higher in WC compared to NC. Being relevant for the empirical strategy and power calculation, the directional hypothesis shall be investigated in further detail. First, let us reconsider that the purpose of this study is to investigate the existence and magnitude of hidden costs of control in the field. Second, a control device has two opposing effects: it may increase work effort by limiting the agents action space (and with it, opportunistic behavior). On the other hand, it may lower work effort through the behavioral mechanism mentioned in

[11] Remember that subjects were randomly assigned to treatment groups.

Behavioral Prediction 1. The question which effects dominates is an empirical one and is of course dependent on the effectiveness of the control device[12]. In order to find hidden costs of control in a between-subject approach, the negative behavioral effect needs to outweigh the beneficial incentive effect. As a consequence, the controlling device must be weak; represented in our study by the low minimum performance requirement in treatment group WC. To sum up, we focus on a weak controlling device where the behavioral effect very likely outweighs the incentive effect[13]. That is why we hypothesize that the implementation of control will not simply lead to statistically different effort levels between the two groups but will *lower* work effort in the treatment group WC.

To conclude, $POST\_OO_i$ is expected to be statistically significantly higher in treatment group WC compared to treatment group NC[14]. Consequently, we will use one-sided tests whenever appropriate.

2.4.2 Secondary Hypotheses

Behavioral Prediction 2. ***Crowding-out of intrinsic motivation.*** *The negative behavioral reaction when assigned to a control environment depends on the extent of intrinsic motivation: highly intrinsically motivated agents exhibit a higher increase in shirking behavior due to the higher potential for crowding-out. Less intrinsically motivated agents exhibit a lower pre-treatment performance and hence, the negative behavioral reaction is limited in magnitude*[15].

Hypothesis 2. ***Heterogeneity: Intrinsic motivation.*** *In interaction with treatment WC, workers playing or regularly watching lacrosse (LACROSSE=1) exhibit higher hidden costs (control-averse reaction). Formally,* $LACROSSE = 1$ *in interaction with the treatment dummy WC is positively correlated with a higher* ΔOO.

Behavioral Prediction 3. ***Negative reciprocity.*** *Agents perceive the implementation of control as a hostile action taken by the principal. The negative behavioral reaction of control-averse agents is a reciprocal action*[16]*. Hence, hidden costs of control work through negative reciprocity: agents with a strong preference for negative reciprocity exhibit a stronger increase in shirking behavior.*

[12] Importantly, we do not dispute that an effective control device is beneficial for the principal.

[13] There is no obvious reason to expect the contrary: inherently, an ineffective controlling device is characterized by its very limited efficiency in enhancing workers effort. The reason is that a ineffective control does not limit the action space of the agent sufficiently enough and hence, workers can still engage in opportunistic behavior.

[14] This directional hypothesis is also backed by previous studies. Literature found almost consistently evidence in favor for the mere existence of hidden costs of control in the presence of an ineffective controlling device. See section 1 Introduction for a literature review.

[15] Inspired by Falk and Kosfeld (2006); Frey (1993).

[16] Based on evidence in literature (Falk & Kosfeld, 2006; Dickinson & Villeval, 2008; Belot & Schröder, 2016).

Hypothesis 3. ***Heterogeneity: Negative Reciprocity.*** *In interaction with treatment WC, a higher preference for negative reciprocity (variable $NEGREC$) is positively correlated with higher hidden costs (ΔOO).*

Behavioral Prediction 4. ***Positive Reciprocity and gift-exchange.*** *No control is perceived as a kind gift and is reciprocated through a lower shirking level. On the other hand, control is an unkind gift and is not positively reciprocated any more (potentially reciprocated negatively, see above). Therefore, workers with a strong preference for positive reciprocity exert more effort in the pre-treatment stage compared to workers with a weak preference for positive reciprocity. As a consequence, the potential for the negative behavioral reaction is higher: workers with a strong preference for positive reciprocity exhibit higher hidden costs.*

Hypothesis 4. ***Heterogeneity: Positive Reciprocity.*** *In interaction with treatment WC, a higher preference for positive reciprocity (variable $POSREC$) is positively correlated with higher hidden costs (variable ΔOO).*

Behavioral Prediction 5. ***Trust.*** *Control-averse subjects perceive the implemented control as a signal of distrust. Therefore, subjects expressing a higher level of general trust are more likely to perceive control as a signal of distrust and thus exhibit higher hidden costs of control*[17]*.*

Hypothesis 5. ***Heterogeneity: Trust.*** *In interaction with treatment WC, a preference for expressed trust (variable $TRUST$) is positively correlated with higher hidden costs of control (variable ΔOO).*

We will also investigate a potential spill-over effect of imposed control to alternative effort dimensions, namely the change of delivered quality measured as the number of errors and the median time elapsed per picture. All variables are summarized in C Appendix: Variable definition on page 22.

2.4.3 Exploratory Research

We collect individual demographic characteristics, fairness considerations and social preferences. We do not form a hypothesis for all of these covariates. However, we will conduct exploratory research with the goal to potentially discover unknown effects. Especially, we will use the different dimensions of the social preferences as a regressor in order to uncover potential correlations between specific social preferences and control-averse behavior. Exploratory variables are clearly labeled as such, refer to the C Appendix: Variable definition on page 22.

[17] Falk and Kosfeld (2006); Sliwka (2007); Masella et al. (2014); Kessler and Leider (2013).

3 Empirical Analysis

3.1 Main analysis

3.1.1 Primary hypothesis: Hypothesis 1

Parametric model

The basic model takes the following form:

$$POST\_OO_i = f(D) \tag{1}$$

where $POST\_OO_i$ is the key outcome of interest, i.e. the count of clicks on the opt-out button in stage 2 under the respective treatment condition. D is a dummy variable indicating the treatment condition. We later gradually include a vector X of demographic control variables, μ representing individual characteristics with regard to intrinsic motivation and fairness and finally δ for individual social preferences. To test hypothesis 1, we apply following linear regression model (OLS), along with non-parametric tests.

$$POST\_OO_i = \beta_0 + \beta_1 D_{iWC} + \epsilon_i \tag{2}$$

whereas:

$POST\_OO_i :=$ endogenous variable representing a measure for work effort respectively **shirking behavior** for each *individual i*

$D_{iWC} :=$ dummy variable indicating treatment condition WC; if $D_{iWC} = 1$ *individual i* is assigned to WC; representing the **exogenous condition**

$\epsilon_i :=$ **error term**, white noise, normally distributed

This basic specification is then extended by including control variables PRE_OO_i (pre-treatment performance), X_{iz} (demographics), μ (intrinsic motivation and fairness) and δ (social preferences). We rerun all regressions with the alternative endogenous variables (e.g. number of errors and time elapsed). This allows us to find potential spillovers: control is imposed on supply (number of clicks allowed for the opt-out button). However, the quality of the work and the time spent is not controlled for at all. As a consequence, we might observe shirking spillovers from the contracted to the non-contracted dimension[18].

With a two sample one-sided t test we assess if the two populations means are equal. We

[18] found by Belot and Schröder (2016). However, we do not have a strong expectation to find such spillovers: in our design, the control device is weak and limits the action space of agents very weakly. Therefore, agents do not need to reciprocate their behavioral reaction in another dimension, because shirking in the contracted dimension is - due to the ineffective control device - itself very cheap.

expect the population mean of WC being higher than of NC.

Non-parametric model

We also test homogeneity of the two samples for stage 2 with a Wilcoxon rank sum (Mann-Whitney U) test to relax the assumption of normal distribution.

3.1.2 Secondary hypotheses

Hypothesis 2 to 5

Parametric model

Hypotheses 2 to 5 investigate potential heterogeneity effects with regard to the control-averse reaction. The basic model takes the following form:

$$\Delta OO_i = f(D) \tag{3}$$

where ΔOO_i is the dependent variable, i.e. a proxy for the difference between workers shirking in stage 2 under the respective treatment condition and workers shirking in stage 1 under a no-control environment. D is a dummy variable indicating the treatment condition. Hypothesis 2 is tested by estimating following specification.

$$\Delta OO_i = \beta_0 + \beta_1 D_{iWC} + \beta_2 LACROSSE_i + \beta_3 D_{iWC} \times LACROSSE_i + \epsilon_i \tag{4}$$

Hypotheses 3 to 5 are essentially estimated with a similar specification, replacing $LACROSSE$ with $NEGREC$, $POSREC$ and $TRUST$ respectively. Control variables are successively included to improve the specification. The following table summarizes the empirical strategy with respect to the basic OLS regressions.

H	Description	OLS specification
1	For the treatment group WC, the number of clicks on the opt-out button will be significantly higher compared to control group NC.	$POST\_OO_i = \beta_0 + \beta_1 D_{iWC} + \epsilon_i$
2	Intrinsic motivated workers exhibit a stronger control-averse reaction.	$\Delta OO_i = \beta_0 + \beta_1 D_{iWC} + \beta_2 LACROSSE_i + \beta_3 D_{iWC} \times LACROSSE_i + \epsilon_i$
3	Workers with a higher preference for negative reciprocity exhibit a stronger control-averse reaction.	$\Delta OO_i = \beta_0 + \beta_1 D_{iWC} + \beta_2 NEGREC_i + \beta_3 D_{iWC} \times NEGREC_i + \epsilon_i$
4	Workers with a higher preference for positive reciprocity exhibit a stronger control-averse reaction.	$\Delta OO_i = \beta_0 + \beta_1 D_{iWC} + \beta_2 POSREC_i + \beta_3 D_{iWC} \times POSREC_i + \epsilon_i$
5	Workers with a higher preference for trust exhibit a stronger control-averse reaction.	$\Delta OO_i = \beta_0 + \beta_1 D_{iWC} + \beta_2 TRUST_i + \beta_3 D_{iWC} \times TRUST_i + \epsilon_i$

Table 2: Empirical strategy summarized: OLS regressions

3.2 Robustness checks

First of all, we will generate another key endogenous variable called $SHIRK$. To make the task naturalistic, we implement images which are indeed unclear and impossible to solve: in each task two out of 20. This variable accounts for that by transforming OO in following way:

$$PRE\_SHIRK_i = \max\{PRE\_OO_i - 2, 0\}$$

number of clicks on the opt-out option in the pre-treatment stage (task 1) which constitute shirking behavior, representing **shirking frequency** for each *individual i*

$$POST\_SHIRK_i = \max\{POST\_OO_i - 2, 0\}$$

number of clicks on the opt-out option in the pre-treatment stage (task 1) which constitute shirking behavior, representing **shirking frequency** for each *individual i*

$$\Delta SHIRK_i = POST\_SHIRK_i - PRE\_SHIRK_i$$

a proxy for the **difference in shirking frequency** between post-treatment and pre-treatment stage, for each *individual i*

Note that the use of this maximum operator is not a linear transformation. Essentially, we create a new endogenous variable where the treatment effect will likely be smaller: imagine a worker who doesn't click the opt-out button in the pre-treatment stage but three times in the treatment stage. His ΔOO is 3 while his $\Delta SHIRK$ is only 1. Therefore, using the $SHIRK$ variable in our regressions demonstrates the robustness of our results.

Second, we will also apply an independent sample t test, a two sample KS test and Wilcoxon rank sum (Mann-Whitney) test to the specifications mentioned in previous section. We include demographic variables as control variables. We especially control for gender differences. Furthermore, we extend the basic specification with number of clicks on the opt-out button squared ($POST\_OO^2$ and ΔOO^2) as control to demonstrate robustness of our results.

Also, we run Logit regressions to estimate the effect on binary outcomes. To do so, agents of treatment group WC are classified into a variable D_{iCA} indicating if workers are control-averse: if agents shirk less in task 1 (no-control environment) than in WC (that is, if $\Delta OOADJ_i > 0$), the endogenous binary variable equals 1, and 0 otherwise. Further, we run logistic regressions to estimate the effect of control on outcomes expressed in proportions. We classify agents according to their reaction to control into one of the following three categories: negative ($\Delta OOADJ_i > 0$, that is control-aversion), neutral ($\Delta OOADJ_i = 0$) and positive ($\Delta OOADJ_i < 0$). Based on ordered logit estimates, we report for the WC experimental condition the predicted probability of falling into one of the three mutually exclusive categories of reaction to control.

If the null of hypothesis 1 cannot be rejected, it may be that the control device is in fact effective and hence, the benefits of control outweigh the indirect negative behavioral cost. To test the mere existence of hidden costs of control, we modify the distribution of shirking behavior[19]. For workers assigned to treatment WC, this modified distribution is expected to be statistically different in the no-control condition (stage 1) and the control condition (stage 2), assessed with a Wilcoxon signed rank test for paired observations. For workers assigned to treatment NC, the modified distribution in stage 1 is expected to be not statistically different than the distribution under stage 2.

[19] We follow F&K's procedure and modify the distribution of transfers respectively in our case the shirking behavior in the no-control condition, that is the pre-treatment condition. Hidden costs are defined if effort in the no-control condition is higher than in the control-condition, that is if $\Delta OO^m = POST\_OO^{WC} - \min\{20 - \underline{x}, PRE\_OO\} > 0$. Any effort strictly lower than $\underline{x}$ is set equal to $\underline{x}$, or in other words, any shirking in the no-control condition being higher than allowed in the condition with a minimum performance requirement is set to the maximum allowed shirking quantity. Maximum shirking when assigned to WC in task two is 8, as subjects are allowed to click maximally 8 times the opt-out button. Maximum shirking in task 1 amounts to 20, as there is no minimum performance requirement. According to F&K, we observe hidden costs if the sum of the ranks of the positive ΔOO^m is sufficiently larger than the sum of the ranks of negative ΔOO^m. To sum up, this procedure neutralizes selfish agents and allows to identify the pure existence of control-averse workers who are responsible for hidden costs of control.

4 Piloting and Power Analysis

4.1 Piloting

Several small pilots were conducted to evaluate the appropriate HIT reward, to assess technical functioning of the Otree environment and to analyze attrition as well as the individual difficulty of the pictures. At the end of July 2018, we conducted a larger pilot ($n = 66$) in two waves to estimate the required sample size. Table 3 presents regression results of the most basic main specification employed for testing Hypothesis 1 (see section 3 Empirical Analysis). The corresponding t test is displayed in Appendix table 3. A first result is that treatment WC indeed leads to higher shirking compared to NC. While workers in NC treatment click on average 3.147 times the "unclear image"-button, workers assigned to WC treatment click it on average 3.75 times, that is 0.603 times more than the NC group. Hence, the treatment effect is relatively large and importantly in the expected direction: weak control lowers workers effort compared to no-control. However, as expected, t-statistics are not significant.

	(1) post_OO
WC	0.603 (.397)
Constant	3.147*** (.276)
Observations	66

standard error in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 3: Regression table pilot study

4.2 Power Analysis

For the calculation of the sample size, we focus on our main hypothesis, which is Hypothesis 1. A one-sided two-sample t test power calculation is computed based on the data presented in the previous section. Figure 3 on page 20 in the Appendix displays a two-sided t test and the relevant summary statistics, figure 4 a one-sided power calculation. Power is set to 0.9, while α is set to 0.05. **The resulting sample size (one-sided) yields 124 subjects per group.**

4.3 Sample size

According to the previous section, we should aim for a total sample of 248 workers, containing data points for both real-effort experimental stages (HIT1 and HIT2) in Part A as well as for the survey module in Part B (HIT3). Workers may not do all individual HITs and may drop out in-between HIT1 and HIT2, and in-between HIT2 and the survey module conducted a week later (HIT3). Dropouts between HIT1 and HIT2 are not harmful as they occur before treatment induction. We expect dropouts between HIT2 and HIT3 to be random and evenly distributed among treatment groups and therefore not relevant to undermine our analysis. In any case, these dropouts will not affect the main hypothesis.

However and importantly, attrition needs to be accounted for to perform a correct calculation of sample size. During piloting, we experienced dropouts between HIT1 and HIT2 amounting to at least 10% and a maximum of 30%. To take a conservative approach, let us assume that we will face 30% attrition.

We have not piloted HIT3 (Part B of the experiment) and thus assume attrition to be of the same magnitude. Therefore, to have a final sample size of 248 subjects, we need to approximately recruit $x \times 0.7^2 = 248$ subjects, that is $\frac{248}{0.7^2} = 506.12$ workers. To conclude, **we will initially recruit 506 workers** by setting the number of individual assignments on AMT for HIT1 to 506, anticipating a final sample of 248 subjects.

5 Practicalities

5.1 Research Team

This research will be conducted by Prof. Dr. Holger Herz and PhD student Christian Zihlmann.

Prof. Dr. Holger Herz
University of Fribourg
Department of Economics
Chair of Industrial Economics
Office G 428
Bd. de Pérolles 90
CH - 1700 Fribourg
holger.herz@unifr.ch

Christian Zihlmann
University of Fribourg
Department of Economics
Chair of Industrial Economics
Office G 412
Bd. de Pérolles 90
CH - 1700 Fribourg
christian.zihlmann@unifr.ch

5.2 Open Science and Timestamp

The principles of the Peer Reviewers' Openness Initiative[20] will be followed. Precisely, we will make our dataset, instructions, material and code to run the statistical analysis publicly available through a trusted third-party repository. Furthermore, to verify and proof existence

[20] see https://opennessinitiative.org/the-initiative/

and non-modification of the dataset as well as the pre-analyis plan, we will timestamp the files on the Bitcoin blockchain[21]. The related hash for verification will be publicly provided.

[21] see `https://opentimestamps.org/`

References

Belot, M., & Schröder, M. (2016). The spillover effects of monitoring: A field experiment. *Management Science*, *62*(1), 37-45. Retrieved from https://doi.org/10.1287/mnsc.2014.2089 doi: 10.1287/mnsc.2014.2089

Boly, A. (2011, May 01). On the incentive effects of monitoring: evidence from the lab and the field. *Experimental Economics*, *14*(2), 241–253. Retrieved from https://doi.org/10.1007/s10683-010-9265-1 doi: 10.1007/s10683-010-9265-1

Burdin, G., Halliday, S., & Landini, F. (2018). The hidden benefits of abstaining from control. *Journal of Economic Behavior & Organization*. Retrieved from http://www.sciencedirect.com/science/article/pii/S0167268117303682 doi: https://doi.org/10.1016/j.jebo.2017.12.018

Chen, D. L., Schonger, M., & Wickens, C. (2016). otree—an open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance*, *9*(Supplement C), 88 - 97. Retrieved from http://www.sciencedirect.com/science/article/pii/S2214635016000101 doi: https://doi.org/10.1016/j.jbef.2015.12.001

Cohn, A., Fehr, E., & Goette, L. (2015). Fair wages and effort provision: Combining evidence from a choice experiment and a field experiment. *Management Science*, *61*(8), 1777-1794. Retrieved from https://doi.org/10.1287/mnsc.2014.1970 doi: 10.1287/mnsc.2014.1970

Dickinson, D., & Villeval, M.-C. (2008). Does monitoring decrease work effort?: The complementarity between agency and crowding-out theories. *Games and Economic Behavior*, *63*(1), 56 - 76. Retrieved from http://www.sciencedirect.com/science/article/pii/S0899825607001364 doi: https://doi.org/10.1016/j.geb.2007.08.004

Falk, A., Becker, A., Dohmen, T., Enke, B., Huffman, D., & Sunde, U. (2018). Global evidence on economic preferences*. *The Quarterly Journal of Economics*, qjy013. Retrieved from http://dx.doi.org/10.1093/qje/qjy013 doi: 10.1093/qje/qjy013

Falk, A., Becker, A., Dohmen, T., Huffman, D., & Sunde, U. (2016). The preference survey module: A validated instrument for measuring risk, time, and social preferences. *IZA Discussion Papers No. 9674*.

Falk, A., Becker, A., Dohmen, T. J., Enke, B., & Huffman, D. (2015). The nature and predictive power of preferences: Global evidence. *IZA Discussion Papers No. 9504*.

Falk, A., & Kosfeld, M. (2006). The hidden costs of control. *The American Economic Review*, *96*(5), 1611-1630. Retrieved from http://www.jstor.org/stable/30034987

Frey, B. S. (1993). Does monitoring increase work effort? the rivalry with trust and loyalty. *Economic Inquiry*, *31*(4), 663–670. Retrieved from http://dx.doi.org/10.1111/

j.1465-7295.1993.tb00897.x doi: 10.1111/j.1465-7295.1993.tb00897.x

Harrison, G. W., & List, J. A. (2004, December). Field experiments. *Journal of Economic Literature*, *42*(4), 1009-1055. Retrieved from http://www.aeaweb.org/articles?id=10.1257/0022051043004577 doi: 10.1257/0022051043004577

Ipeirotis, P. G. (2010, December). Analyzing the amazon mechanical turk marketplace. *XRDS*, *17*(2), 16–21. Retrieved from http://doi.acm.org/10.1145/1869086.1869094 doi: 10.1145/1869086.1869094

Kessler, J., & Leider, S. (2012). Norms and contracting. *Management Science*, *58*(1), 62-77. Retrieved from https://doi.org/10.1287/mnsc.1110.1341 doi: 10.1287/mnsc.1110.1341

Kessler, J., & Leider, S. (2013). *Finding the cost of control* (CESifo Working Paper: Bavioural Economics No. 4188). Munich. Retrieved from http://hdl.handle.net/10419/71266

Kessler, J., & Leider, S. (2016). Procedural fairness and the cost of control. *The Journal of Law, Economics, and Organization*, *32*(4), 685-718. Retrieved from +http://dx.doi.org/10.1093/jleo/eww009 doi: 10.1093/jleo/eww009

Masella, P., Meier, S., & Zahn, P. (2014). Incentives and group identity. *Games and Economic Behavior*, *86*(Supplement C), 12 - 25. Retrieved from http://www.sciencedirect.com/science/article/pii/S089982561400044X doi: https://doi.org/10.1016/j.geb.2014.02.013

Nagin, D. S., Rebitzer, J. B., Sanders, S., & Taylor, L. J. (2002). Monitoring, motivation, and management: The determinants of opportunistic behavior in a field experiment. *The American Economic Review*, *92*(4), 850-873. Retrieved from http://www.jstor.org/stable/3083284

Riener, G., & Wiederhold, S. (2016). Team building and hidden costs of control. *Journal of Economic Behavior & Organization*, *123*(Supplement C), 1 - 18. Retrieved from http://www.sciencedirect.com/science/article/pii/S0167268115003327 doi: https://doi.org/10.1016/j.jebo.2015.12.008

Schmelz, K., & Ziegelmeyer, A. (2015). *Social distance and control aversion : Evidence from the internet and the laboratory* (Tech. Rep. No. 100). Konstanz: Thurgauer Wirtschaftsinstitut. Retrieved from http://nbn-resolving.de/urn:nbn:de:bsz:352-0-322085

Schnedler, W., & Vadovic, R. (2011). Legitimacy of control. *Journal of Economics & Management Strategy*, *20*(4), 985–1009. Retrieved from http://dx.doi.org/10.1111/j.1530-9134.2011.00315.x doi: 10.1111/j.1530-9134.2011.00315.x

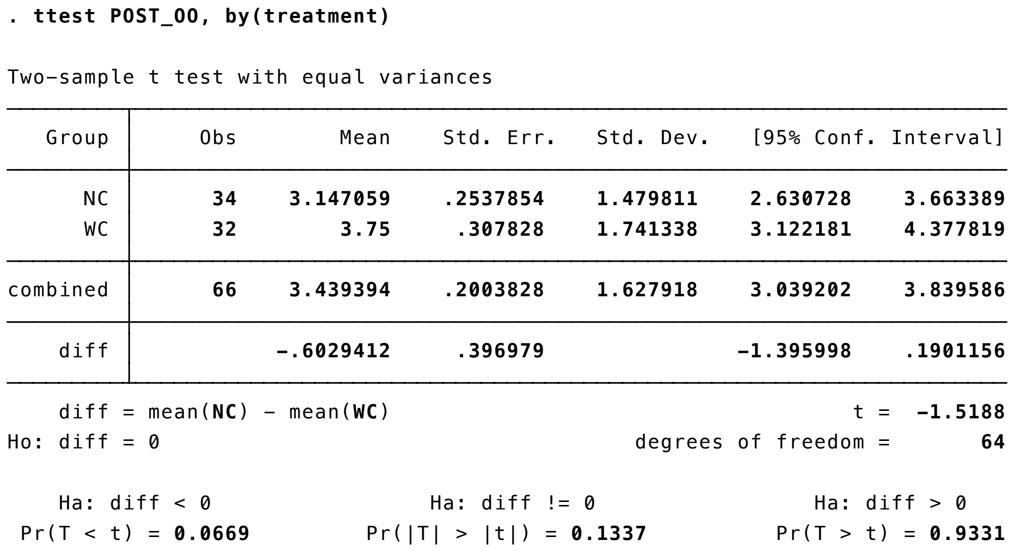
Sliwka, D. (2007). Trust as a signal of a social norm and the hidden costs of incentive schemes. *The American Economic Review*, *97*(3), 999-1012. Retrieved from http://www.jstor

.org/stable/30035032

Ziegelmeyer, A., Schmelz, K., & Ploner, M. (2012, Jun 01). Hidden costs of control: four repetitions and an extension. *Experimental Economics*, *15*(2), 323–340. Retrieved from https://doi.org/10.1007/s10683-011-9302-8 doi: 10.1007/s10683-011-9302-8

Appendices

A Appendix: Additional tables

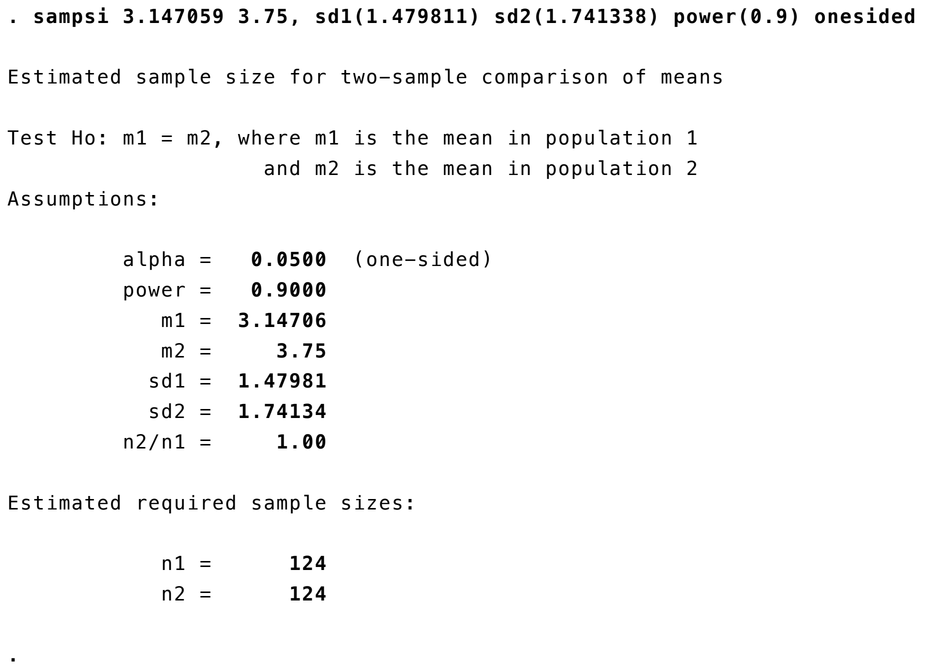


```
. ttest POST_OO, by(treatment)

Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
      NC |      34    3.147059    .2537854    1.479811    2.630728    3.663389
      WC |      32        3.75     .307828    1.741338    3.122181    4.377819
---------+--------------------------------------------------------------------
combined |      66    3.439394    .2003828    1.627918    3.039202    3.839586
---------+--------------------------------------------------------------------
    diff |           -.6029412     .396979               -1.395998    .1901156
------------------------------------------------------------------------------
    diff = mean(NC) - mean(WC)                                    t =  -1.5188
Ho: diff = 0                                     degrees of freedom =       64

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0669         Pr(|T| > |t|) = 0.1337          Pr(T > t) = 0.9331
```

Figure 3: Two sample two tailed t test pilot study



```
. sampsi 3.147059 3.75, sd1(1.479811) sd2(1.741338) power(0.9) onesided

Estimated sample size for two-sample comparison of means

Test Ho: m1 = m2, where m1 is the mean in population 1
                    and m2 is the mean in population 2
Assumptions:

         alpha =   0.0500  (one-sided)
         power =   0.9000
            m1 =  3.14706
            m2 =     3.75
           sd1 =  1.47981
           sd2 =  1.74134
         n2/n1 =     1.00

Estimated required sample sizes:

            n1 =      124
            n2 =      124

.
```

Figure 4: Sample size calculation one-tailed based on pilot study

B Appendix: Formal hypotheses

The research hypotheses are in bold (expected outcome).

Hypothesis 1

OLS:

$H_0 : \beta_1 = 0$

$\mathbf{H_1} : \beta_{\mathbf{1}} > \mathbf{0}$

Student's t-test (one-sided):

$H_0 : \mu_{NC} - \mu_{WC} = 0$

$\mathbf{H_1} : \mu_{\mathbf{NC}} - \mu_{\mathbf{WC}} < \mathbf{0}$

Wilcoxon Mann-Whitney U test on $POST\_OO$.

$H_0 : \alpha \leq 0$

$\mathbf{H_1} : \alpha > \mathbf{0}$

Hypothesis 2 to 5

OLS:

$H_0 : \beta_3 = 0$

$\mathbf{H_1} : \beta_{\mathbf{3}} > \mathbf{0}$

C Appendix: Variable definition

Variable (stata name)	Mined in task	Group	Description	Properties	Purpose	Hypothesis
PRE_OO	1	Y	number of clicks on "Unclear image"-button	discrete: min 0, max 20	auxiliary variable	none
PRE_ERRORS	1	Y	sum of total errors for all 20 pictures (max 5 per picture)	discrete: min 0, max 100	auxiliary variable	none
PRE_ERRORSB	1	Y	sum of wrongly solved pictures, regardless of how many errors per picture	discrete: min 0, max 20	auxiliary variable	none
PRE_TIME	1	Y	sum of total time elapsed to do 20 pictures	continuous	auxiliary variable	none
PRE_MTIME	1	Y	median time elapsed to solve one picture (out of all 20)	continuous	auxiliary variable	none
PRE_FOCUS	1	Y	total time elapsed to do 20 pictures, only when worker is active in the browser window	continuous	auxiliary variable	none
PRE_OIM	1	CV	number of clicks to open images in large scale	discrete: min 0, max 20	CV	CV
PRE_OINS	1	CV	number of clicks on the "Show instructions" button	discrete: min 0, max unlimited	CV	CV
PRE_OOMO	constructed	Y	Neutralization of selfish agents: $\min\{20-\underline{x}, PRE\_OO\}$	discrete: min 0, max 8	auxiliary variable	none
PRE_SHIRK	constructed	Y	max(PRE_OO -2, 0). PRE_OO adjusted for the two unclear picture included =Count of images shirked in task 1	discrete: min 0, max 18	auxiliary variable	none
PRE_SHIRKMO	constructed	Y	Neutralization of selfish agents: $\min\{20-2-\underline{x}, PRE\_SHIRK\}$	continuous: min 0, max 6	auxiliary variable	none
POST_OO	**2**	**Y**	**number of clicks on "Unclear image"-button**	**discrete: min 0, max 20**	**endogenous**	**Hypothesis 1**
POST_ERRORS	2	Y	sum of total errors for all 20 pictures (max. 5 per picture)	discrete: min 0, max 100	alternative Y (spill-over effect)	Hypothesis 1
POST_ERRORSB	2	Y	sum of wrongly solved pictures, regardless of how many errors per picture	discrete: min 0, max 20	alternative Y (spill-over effect)	Hypothesis 1
POST_TIME	2	Y	sum of total time elapsed to do 20 pictures	continuous	auxiliary variable	none
POST_MTIME	2	Y	median time elapsed to solve one picture (out of all 20)	continuous	alternative Y (spill-over effect)	Hypothesis 1
POST_FOCUS	2	Y	total time elapsed to do 20 pictures, active window	continuous	alternative Y (spill-over effect)	Hypothesis 1
POST_OIM	2	CV	number of clicks to open images in large scale	discrete: min 0, max 20	CV	CV
POST_OINS	2	CV	number of clicks on the "Show instructions" button	discrete: min 0, max unlimited	CV	CV
POST_OOCOM	constructed	Y	min(POST_OO, 8). Variable taking into account subjects who do not comply with the MPR: max. clicks allowed is 8, so that all comply.	discrete: min 0, max 8	auxiliary variable	none
POST_SHIRK	constructed	Y	max(POST_OO -2, 0). POST_OO reduced by the two unclear pictures included =Count of images shirked in task 2	discrete: min 0, max 18	endogenous (robustness)	Hypothesis 1
POST_SHIRKCOM	constructed	Y	max(max(POST_OO -2, 0), 6). Variable taking into account subjects who do not comply with the MPR: max. shirking set to 6 so that all comply.	discrete: min 0, max 6	auxiliary	none
d_OO	**constructed**	**Y**	**POST_OO - PRE_OO**	**discrete: min -20, max 20**	**endogenous**	**Hypothesis 2,3,4,5**
d_OOBASE	constructed	Y	mean of: POST_OO - PRE_OO; for subjects assigned to NC only	discrete: min -20, max 20	auxiliary	none
d_OOBASEM	constructed	Y	mode of: POST_OO - PRE_OO; for subjects assigned to NC only	discrete: min -20, max 20	auxiliary	none
d_OOADJ	constructed	Y	POST_OO - PRE_OO - d_OOBASE	discrete: min -40, max 40	endogenous	Logit and Probit robustness tests
d_OOADJMODE	constructed	Y	POST_OO - PRE_OO - d_OOBASEM	discrete: min -40, max 40	endogenous	Logit and Probit robustness tests
d_SHIRK	constructed	Y	POST_SHIRK - PRE_SHIRK	discrete: min -18, max 18	endogenous (robustness)	Hypothesis 2,3,4,5
d_SHIRKBASE	constructed	Y	mean of: POST_SHIRK - PRE_SHIRK; for subjects assigned to NC only	discrete: min -18, max 18	auxiliary	none
d_SHIRKBASEM	constructed	Y	mode of: POST_SHIRK - PRE_SHIRK; for subjects assigned to NC only	discrete: min -18, max 18	auxiliary	none
d_SHIRKADJ	constructed	Y	POST_SHIRK - PRE_SHIRK - d_SHIRKBASE	discrete: min -36, max 36	auxiliary variable	none
d_SHIRKADJMODE	constructed	Y	POST_SHIRK - PRE_SHIRK - d_SHIRKBASEM	discrete: min -36, max 36	auxiliary variable	none
			a "R" at the end of each variable (where possible) denotes its relative value			

Table 4: Summary Of Variables, group Y

Variable (stata name)	Mined in task	Group	Description	Properties	Purpose	Hypothesis
d_ERRORS	constructed	Y	POST_ERRORS - PRE_ERRORS	discrete: min -100, max 100	alternative Y (spill-over effect)	Hypothesis 2,3,4,5
d_ERRORSB	constructed	Y	POST_ERRORSB - PRE_ERRORSB	discrete: min -20, max 20	alternative Y (spill-over effect)	Hypothesis 2,3,4,5
d_TIMEM	constructed	Y	POST_TIMEM-PRE_TIMEM	continuous	alternative Y (spill-over effect)	Hypothesis 2,3,4,5
d_FOCUS	constructed	Y	POST_FOCUS - PRE_FOCUS	continuous	endogenous (robustness)	Hypothesis 1
CA	constructed	Y	Dummy, 1 if $d\_OOADJ > 0$	Categorical (dichotomous)	endogenous (robustness)	Probit Hyp. 1
TYPE	constructed	Y	Control-averse: CA if $d\_OOADJ > 0$; Neutral: NE if $d\_OOADJ = 0$; Opportunistic: OP if $d\_OOADJ < 0$	Categorical	endogenous (robustness)	Logit Hyp. 1
treatment1	2	D	Treatment dummy "string" variables. NC = no control, WC= weak control.	categorical	auxiliary variable	none
treatment	constructed	D	Treatment dummy encoded as numeric variable	Categorical (dichotomous)	exogenous variation	exogenous variation
LACROSSE	2	μ	Elicitation of a workers familiarity with the sport of lacrosse: Do you play or regularly watch Lacrosse? Proxy for intrinsic motivation	Categorical (dichotomous)	Heterogeneity analysis	Hypothesis 2
FAIRL	2	μ	Measures individual fairness perceptions with regard to the HIT reward on a 7-point likert scale: On a scale from 1 to 7, how fair do you consider the reward we pay you for this HIT? (1 very unfair, 7 very fair)	Likert: 1 to 7	exploratory	[22]
FAIRQ	2	μ	Measures individual fairness perceptions with regard to the HIT reward quantitatively: What reward would be appropriate for doing your work? I consider a HIT reward of x.xx USD (enter value below) to be appropriate.	discrete: min 0, max 50	exploratory	see above
DEVICE	2	μ	What device are you currently using?	Categorical (nominal), 4	CV	CV
MUSIC	2	μ	Elicitation of a workers familiarity with music (placebo): Do you play a musical instrument? Placebo for LACROSSE	Categorical (dichotomous)	CV	CV
BOOK	2	μ	Elicitation of a workers familiarity with books (placebo): Do you enjoy reading books? Placebo for LACROSSE	Categorical (dichotomous)	CV	CV
SEX	3	X	Demographics: Gender.	Categorical (dichotomous)	CV	CV
AGE	3	X	Demographics: Age. Categories: < 20, $21 - 29$, $30 - 39$, $40 - 49$, $50 - 59$, > 60	Categorical (ordinal), 6	CV	CV
EDU	3	X	Demographics: Education. Categories: Less than high school degree, High school degree or equivalent (e.g. GED), Some college but no degree, Associate degree, Bachelor degree, Graduate degree (e.g. Master degree)	Categorical (ordinal), 6	CV	CV
RACE	3	X	Demographics: Race. Categories: Black or African American, American Indian or Alaskan Native, White or Caucasian, Hispanic or Spanish or Latino, Asian or Pacific Islander, Other or none of the listed, Native American	Categorical (nominal), 7	CV	CV
REL	3	X	Demographics: Religion. Categories: No religion / Atheism, Other, Judaism, Christianity, Buddhism, Hinduism, Native American, Islam	Categorical (nominal), 8	CV	CV
WEEKH	3	X	Demographics: Mturk weekly work hours. 0 to 9, 10 to 19, 20 to 29, 30 to 39, 40 and more	Categorical (ordinal), 5	CV	CV
REAS	3	X	Demographics: Reason for Mturk participation. Learn new skills, Make money, Kill time, Other, Have fun	Categorical (nominal), 5	exploratory	Reasons indicating intrinsic motivation, such as "learn new skills" and "have fun" might display the same pattern as formulated in Hypothesis 2

Table 5: Summary Of All Variables cont'd, groups Y, D, μ and X

[22] The purpose is to investigate the heterogeneous treatment effect in interaction with the treatment dummy. Two contradicting hypotheses: 1) High FAIRL leads (in interaction with WC) to lower effort (trust destroyed -> disappointment effect see Masella et al. (2014) and Kessler and Leider (2012). Agents expect principal to not control, potential for neg. reaction higher.) 2) High FAIRL leads (in interaction with WC) to higher effort (control is perceived as legitimate see Kessler and Leider (2016)) 3) low FAIRL (in interaction with WC) leads to lower effort (negative reciprocity and procedural fairness) 4) low FAIRL (in interaction with WC) leads to higher effort (potential very low for hidden costs to arise as effort in NC anyway low. Control is beneficial when week social norm according to Kessler and Leider (2013).)

Variable (stata name)	Mined in task	Group	Description	Properties	Purpose	Hypothesis
arisk1	3	δ	Risk. Self-assessment: Willingness to take risks in general	Likert: 0 to 10	auxiliary variable	-
atime1	3	δ	Patience. Self-assessment: Willingness to wait. How willing are you to give up something that is beneficial for you today in order to benefit more from that in the future?	Likert: 0 to 10	auxiliary variable	-
anegrec1	3	δ	Negative reciprocity. Self-assessment: Willingness to punish unfair behavior towards self. How willing are you to punish someone who treats you unfairly, even if there may be costs for you?	Likert: 0 to 10	auxiliary variable	-
anegrec2	3	δ	Negative reciprocity. Self-assessment: Willingness to punish unfair behavior towards others. How willing are you to punish someone who treats others unfairly, even if there may be costs for you?	Likert: 0 to 10	auxiliary variable	-
aaltruism1	3	δ	Altruism. Self-assessment: Willingness to give to good causes. How willing are you to give to good causes without expecting anything in return?	Likert: 0 to 10	auxiliary variable	-
aposr1	3	δ	Positive reciprocity. Self-assessment: Willingness to return a favor. When someone does me a favor I am willing to return it.	Likert: 0 to 10	auxiliary variable	-
anegrec3	3	δ	Negative reciprocity. Self-assessment: Willingness to take revenge. If I am treated very unjustly, I will take revenge at the first occasion, even if there is a cost to do so.	Likert: 0 to 10	auxiliary variable	-
atrust	3	δ	Trust. Self-assessment: People have only the best intentions. I assume that people only have best intentions.	Likert: 0 to 10	auxiliary variable	-
arisk2	3	δ	Risk taking. Lottery choice sequence using staircase method	Categorical (ordinal), 32	auxiliary variable	-
atime2	3	δ	Patience. Inter-temporal choice sequence using staircase method.	Categorical (ordinal), 32	auxiliary variable	-
aposr2	3	δ	Positive reciprocity. Gift in exchange for help.	Categorical (ordinal), 7	auxiliary variable	-
aaltruism2	3	δ	Altruism. Donation decision: Imagine the following situation: Today you unexpectedly received 1,000 USD. How much of this amount would you donate to a good cause?	Discrete, min 0 max 1000	auxiliary variable	-
			a "z" instead an "a" in front of each of the above mentioned a-variable indicates the z-score (computed for each survey item at the individual level).			
PATIENCE	constructed	δ	0.7115185 x ztime2 + 0.2884815 x ztime1	Continuous (z-score)	exploratory	No clear hypothesis. A higher patience could be attributed with a higher effort level .
RISK	constructed	δ	0.4729985 x zrisk2 + 0.5270015 x zrisk1	Continuous (z-score)	exploratory	No clear hypothesis. A higher willingness to take risk could be attributed with a shirking level.
POSREC	constructed	δ	0.4847038 x zposr1 + 0.5152962 x zposr2	Continuous (z-score)	Heterogeneity analysis.	Hypothesis 4
NEGREC	constructed	δ	(0.6261938/2) x znegrec1 + (0.6261938/2) x znegrec2 + 0.3738062 x znecrec3	Continuous (z-score)	Heterogeneity analysis.	Hypothesis 3
ALTRUISM	constructed	δ	0.6350048 x zaltruism1 + 0.3649952 x zaltruism2	Continuous (z-score)	exploratory	A lower score on altruism may be positively correlated with lower effort as selfish individuals tend to shirk more often.
TRUST	constructed	δ	ztrust	Continuous (z-score)	Heterogeneity analysis.	Hypothesis 5

Table 6: Summary Of All Variables cont'd, group δ social preferences