UNIVERSITÉ DE FRIBOURG
UNIVERSITÄT FREIBURG

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

University of Fribourg

05.2 Pre-analysis plan: Replication

08.03.2020

Replication of: 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 experiment is to replicate the findings from our first experimental run, which was pre-registered in October 2018 and conducted in December 2018. This pre-analysis plan shall briefly outline the replication procedure.

2 Research Strategy

2.1 Experimental Design

This study replicates the first run of the experiment that was conducted in December 2018 (refer to the Pre-Analysis Plan dated 22 October 2018). The experimental design of this replication is without any modification compared to the first run. In the following, we briefly recapitulate the study design.

The natural field experiment is conducted on an online crowdsourcing labor market, namely Amazon Mechanical Turk ("AMT"). Workers are not aware that they are participating in an experiment and engage in a visual search task: extracting and categorizing information from a picture. Concretely, we present workers with pictures from game-play situations of a lacrosse match and ask them to extract five pieces of information from each picture.

Before starting to work on an image, the worker first needs to decide whether the image is readable. This is the case if all requested information is visible ("Clear image, all info visible"-button). Workers can also decide to opt-out ("Unclear image, not all info visible"-button): This button is the truthful response if workers cannot solve a picture; e.g. if the picture is blurry or the requested information is not readable. Such an opt-out option is commonly used on AMT and in our setting, it allows for cheap shirking: If a worker reports an image as unclear, then he skips it and moved on to the next image. Workers are instructed that it is well possible for some images to be blurry or unreadable. Importantly, reporting those is not going to reduce their payment. In both stages, we included two unreadable images. Consequently, skipping more than two pictures constitutes misbehavior.

The experiment consists of two parts, a pre-treatment stage and a treatment stage, also referred to as task 1 and task 2. In each stage, workers are tasked with categorizing a set of 20 pictures. In both stages, workers receive a flat fee.

In the pre-treatment stage, all workers are subject to an environment where monitoring is absent: workers are paid in any case. Workers can potentially click the opt-out button 20 times without transcribing a single image and still receive the full reward.

Once workers have completed the pre-treatment stage, they are offered the opportunity to do a different set of 20 pictures in another HIT. This is the treatment stage where the contract of workers is varied. The control group (henceforth: "Baseline") receives the same contract as before and is not subject to any monitoring mechanism. For the treatment group (henceforth: "Monitored"), a control mechanism in the form of a minimum performance requirement is implemented: Workers are allowed to click the opt-out option maximally 8 out of 20 times. Workers are automatically paid if they do not exceed this threshold - if they do, they are not eligible to receive the reward. Note that we do not impose any control on work quality, a fact known to workers.

2.2 Main outcome variables

Table 1 provides an overview about the key outcome variables collected in the real effort task. In essence, we observe work input as well as work output, two proxies for unobserved effort.

Table 1: Key Outcome Variables

Variable name	Dimension	Description	Properties
OUTPUT	Work output / Performance	Number of correctly transcribed pictures, total work output (=20-SKIP-ERRORS).	min:0 max:20
SKIP	Misbehavior	Number of skipped readable images.	min:0 max:18
ERRORS	Misbehavior	Number of transcribed images that contain an error.	min:0 max:20
INPUT	Work input	Time elapsed to complete the task.	continuous

Note: The prefix *pre* indicates task 1 (pre-treatment stage), the prefix *post* indicates task 2 (treatment stage).

Work output is multi-dimensional and composed of two sub-dimensions that reflect the two different ways workers can engage in misbehavior:

SKIP. Through a click on the "Unclear image"-button, workers can declare readable and clear images as unclear, resulting in skipping these readable images. Workers thus avoid to work on these as unreadable declared images.

ERRORS. Another way of shirking is to correctly declare readable images as clear images, but then making errors when sloppily transcribing them. This misbehavior results in images transcribed with errors. The two misbehavior opportunities are likely close substitutes.

OUTPUT. Work output incorporates both misbehavior (skipping readable images and making errors) by measuring the total number of correctly transcribed images - the outcome a principal is ultimately interested in.

INPUT. Furthermore, we track workers activity and thus provide process data: Work input represents time on task, a valid proxy for work effort or attention (Gabaix, 2019). We use focus time elapsed to measure time on task, a novel measure collected with `otree_tools` (Chapkovski

& Zihlmann, 2019).

2.3 Hypotheses

2.3.1 Main analysis

Hypothesis 1

We specify the following two hypotheses, again following the Pre-Analysis Plan dated 22 October 2018.

Hypothesis 1. ***Average Treatment Effect.*** *Workers reduce performance when monitored.*

For Hypothesis 1 to be supportable, workers need to decrease their performance when the employer imposes a monitoring device, so when they are treated. *We hypothesize that monitored (treated) workers provide lower work OUTPUT compared to non-monitored workers (control).* Note that shirking in the monitored sub-dimension might be costlier than doing so in the non-monitored sub-dimension. Thus, the behavioral effect might not occur in the monitored (SKIP) sub-dimension, but rather in the non-monitored sub-dimension (ERRORS). We hypothesize that treated workers reduce performance compared to non-treated workers by shirking in the non-monitored sub-dimension, i.e. by making more ERRORS.

Hypothesis 2

Hypothesis 2. ***Crowding-out of intrinsic motivation. Heterogeneous treatment effect.*** *Intrinsically motivated workers reduce performance when monitored.*

Hypothesis 2 accounts for and tests the hypothesis that workers behavioral reaction in response to the monitoring device will likely be heterogeneous. We expect intrinsically motivated workers to reduce their work effort when monitored. We will identify intrinsic motivation through pre-treatment work INPUT, i.e. focus time elapsed, as as a proxy for intrinsic motivation. *We hypothesize that monitored, motivated workers provide lower work OUTPUT compared to motivated non-treated workers..* Again, this behavioral reaction is expected to happen in the non-monitored misbehavior sub-dimension, that is by reducing work output through making more ERRORS.

2.3.2 Secondary Analysis

Hypothesis 3

In this experiment, workers are tasked with transcribing 20 different images with a varying difficulty. We expect workers who are crowded-out by the monitoring mechanism to shirk especially among difficult pictures that require more effort. In other words, the crowding-out is expected to happen among the harder picture categories: workers reduce their work effort where it is cheapest for them to do so, namely among pictures where they can avoid the high cost (high effort).

Hypothesis 3. ***Crowding-out among the difficult sub-tasks.*** *Workers who exhibit a motivational crowding-out are expected to do so by reducing their work OUTPUT among the difficult picture categories.*

3 Empirical Analysis

3.1 Main analysis

3.1.1 Hypothesis 1: Average treatment effect.

The basic model takes the following form:

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

where $POST\_Y_i$ represents the observed performance or the observed misbehavior, as specified in the previous section (variables OUTPUT, SKIP, ERRORS). D is a dummy variable indicating the treatment condition. We test hypothesis 1 by analysing the central tendency with Welch's t-test along with Mann-Whitney-U tests. We will also apply Epps-Singleton tests to compare the distributions. As robustness, we will perform the test also on ΔY_i, that is the change score (post minus pre-treatment measurement) of the respective outcome variable under investigation.

3.1.2 Hypothesis 2: Heterogeneous treatment effect

Hypothesis 2 posits that motivated workers exhibit hidden costs of control due to motivational crowding-out.

Regression approach

Hypothesis 2 is tested with following regression specification:

$$POST\_OUTPUT_i = \beta_0 + \beta_1 D_i + \beta_2 PRE\_INPUT_i + \beta_3 D_i \times PRE\_INPUT_i + \epsilon_i \tag{2}$$

where $POST\_OUTPUT_i$ is the dependent variable, i.e. the proxy for worker's performance. We will run OLS regressions with robust standard errors since the data is believed to be approximately normally distributed. *For our hypothesis to be supportable, β_3 should be statistically significantly lower than zero.*

We will also run the regression specification with the outcome variables ERRORS and SKIP, in order to see in which sub-dimension motivated workers shirk. As outlined previously, we expect workers to shirk in the non-monitored sub-dimension, that is by making more ERRORS. Thus, we hypothesize that β_3 should be statistically significantly greater than zero when ERRORS is the outcome variable.[1].

Median split approach

In a second step, we will classify workers into two types based on a median split of pre-treatment work INPUT: little and highly motivated workers. The two groups will reduce statistical power, but facilitate interpretation of the results. We will do the analysis with the difference between post-treatment and pre-treatment stage (change scores) and as robustness test with the regressor variable approach[2]. We will analyze the proxies for work performance and misbehavior (OUTPUT, SKIP, ERROR) as well as work INPUT. We will test for the difference-of-means between treated vs. control separately for little motivated workers and for highly motivated workers by applying Welch's t-test and Mann-Whitney-U tests.

Robustness

As a main robustness test for hypothesis 2, we will perform the same analysis with PRE_OUTPUT instead of PRE_INPUT as a proxy for identifying worker's intrinsic motivation. Note that OUTPUT is discrete (min 0, max 20), while INPUT is continuous. This will reduce statistical power in the regression approach outlined in the first paragraph. For the median split approach, we might not be able to split our sample at exactly the median, which also likely reduces statistical power.

[1] For SKIP, we expect highly right-skewed data. Thus, in addition to OLS, we will also apply a poisson regression for SKIP. For SKIP, we do not expect a interaction effect, since this is the monitored dimension - therefore, β_3 should be statistically not significantly different from zero.

[2] A pre-treatment post-treatment control group design can be analyzed i) with the post-treatment measurement as dependent variable and the pre-treatment measurement as a co-variate (commonly referred to as ANCOVA) or ii) with the differences between post-treatment and pre-treatment measurements as dependent variable (commonly referred to as CHANGE SCORE), see e.g. Allison (1990) and Lord (1967). If treatment assignment is random - which it is in our case - both methods are unbiased (Breukelen, 2006; Wright, 2006). Actually, using both methods is proposed to be a good practice (Allison, 1990).

3.2 Secondary Analysis

3.2.1 Hypothesis 3: Crowding-out among hard sub-tasks

In this experiment, workers are tasked with transcribing 20 images. Two of them are unreadable and thus impossible to transcribe. By making use of the panel data structure, we will classify the remaining 18 images into three categories of each six pictures, based on their difficulty (easy, medium, hard). The categorization is objectively based on the Baseline group performance - which was not subject to monitoring - of the December 2018 experiment[3]. We will then run following linear regression:

$$POST\_OUTPUT_{it} = D_i + CATEGORY_{it} + D_i \times CATEGORY_{it} + \epsilon_{it} \tag{3}$$

For our hypothesis to be supportable, the coefficient of the interaction term for the category "hard" must be statistically significantly lower than zero. In light of Hypothesis 2, we will also run the same regression for motivated and non-motivated workers separately. We will also test the hypothesis non-parametrically: We test for differences among the treatment groups by applying Mann-Whitney-U tests to the mean percentage of work OUTPUT separately for each of the three categories. Subsequently, we will repeat this test with the median splitted sample, that is for motivated and non-motivated workers separately, again in light of Hypothesis 2.

4 Sampling and Procedures

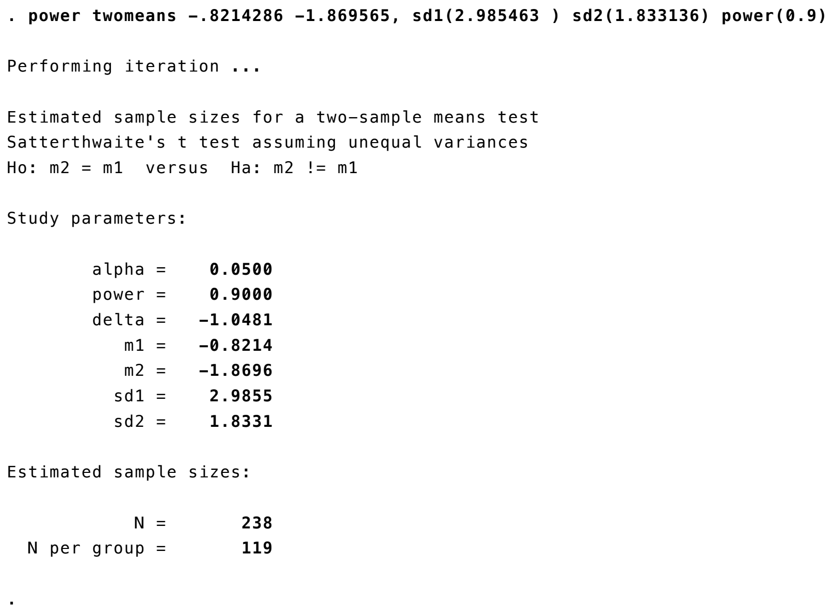
4.1 Power Analysis

The following power analysis is based on the full run of the December 2018 experiment, which was also pre-registered and which we want to replicate. The sample size is calculated based on Hypothesis 2, the median split approach, which requires the most statistical power of the hypotheses covered in the main analysis.

In our first experiment, motivated workers reduced work OUTPUT when treated compared to non-treated motivated workers. This effect is of key interest. That is why the power ana lysis is based on a comparison of motivated workers (treated vs control) with regard to the outcome variable OUTPUT. We aim for a power of 90%. Using the means and standard deviations from the December 2018 experiment (refer to Figure 1 for the stata output) yields a sample size of 119 per group, or in total, 238 motivated workers, divided into treated vs non-treated. Note that half of the workers will be classified as little motivated, so the total sample size should

[3] Concretely, the easy category consists of pictures 15,27,40,54,78,72; medium: 25,32,67,79,33,1) and hard: 68,13,5,19,74,95.

yield 476 workers.



```
. power twomeans -.8214286 -1.869565, sd1(2.985463 ) sd2(1.833136) power(0.9)

Performing iteration ...

Estimated sample sizes for a two-sample means test
Satterthwaite's t test assuming unequal variances
Ho: m2 = m1  versus  Ha: m2 != m1

Study parameters:

        alpha =    0.0500
        power =    0.9000
        delta =   -1.0481
           m1 =   -0.8214
           m2 =   -1.8696
          sd1 =    2.9855
          sd2 =    1.8331

Estimated sample sizes:

            N =       238
  N per group =       119

.
```

Figure 1: Power Calculation

As just outlined, we should aim for a **total sample size of 476 workers**, containing data points for both real-effort experimental stages (HIT1 and HIT2). Workers may not do both HITs and may drop out in-between HIT1 and HIT2. Dropouts between HIT1 and HIT2 are not harmful for statistical inference as they occur before treatment induction. However, we need to account for it for the calculation of sample size. 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. Therefore, to have a final sample size of 476 subjects, we need to approximately recruit $\frac{476}{0.7} = 680$ workers. In short, we will initially recruit 680 workers by setting the number of individual assignments on AMT for HIT1 to 680, anticipating a final sample of 476 subjects.

4.2 Sample Characteristics

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 e.g. Master's qualifications. Workers will be randomly assigned to the treatment and control group, constituting the exogenous variation in this study. The sample will consist of the workers who completed HIT2. We will exclude from our sample all observations for which we do not have all data points - e.g. workers who only complete HIT1 but do not proceed to HIT2 or workers with missing data points such as missing focus time. In general, apart from such cases, all observations will be in the final sample.

4.3 Procedure

The experiment will be conducted from 9 to 13 March 2020. Each day at the same time, we will post the same number of HIT assignments on AMT.

References

Allison, P. D. (1990). Change scores as dependent variables in regression analysis. *Sociological Methodology*, *20*, 93–114. Retrieved from http://www.jstor.org/stable/271083

Breukelen, G. J. V. (2006). Ancova versus change from baseline had more power in randomized studies and more bias in nonrandomized studies. *Journal of Clinical Epidemiology*, *59*(9), 920 - 925. Retrieved from http://www.sciencedirect.com/science/article/pii/S0895435606000813 doi: https://doi.org/10.1016/j.jclinepi.2006.02.007

Chapkovski, P., & Zihlmann, C. (2019). Introducing otree_tools: A powerful package to provide process data for attention, multitasking behavior and effort through tracking focus. *Journal of Behavioral and Experimental Finance*, *23*, 75 - 83. Retrieved from http://www.sciencedirect.com/science/article/pii/S2214635018302119 doi: https://doi.org/10.1016/j.jbef.2019.04.010

Gabaix, X. (2019). Chapter 4 - behavioral inattention. In B. D. Bernheim, S. DellaVigna, & D. Laibson (Eds.), *Handbook of behavioral economics - foundations and applications 2* (Vol. 2, p. 261 - 343). North-Holland. Retrieved from http://www.sciencedirect.com/science/article/pii/S2352239918300216 doi: https://doi.org/10.1016/bs.hesbe.2018.11.001

Lord, F. M. (1967). A paradox in the interpretation of group comparisons. *Psychological bulletin*, *68*(5), 304.

Wright, D. B. (2006). Comparing groups in a before–after design: When t test and ancova produce different results. *British Journal of Educational Psychology*, *76*(3), 663-675. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1348/000709905X52210 doi: 10.1348/000709905X52210