CESifo WORKING PAPERS

8890
2021

Original Version: February 2021
This Version: May 2022

Adverse Effects of Control: Evidence from a Field Experiment

Holger Herz, Christian Zihlmann

CESifo

Impressum:

CESifo Working Papers
ISSN 2364-1428 (electronic version)
Publisher and distributor: Munich Society for the Promotion of Economic Research - CESifo GmbH
The international platform of Ludwigs-Maximilians University's Center for Economic Studies and the ifo Institute
Poschingerstr. 5, 81679 Munich, Germany
Telephone +49 (0)89 2180-2740, Telefax +49 (0)89 2180-17845, email office@cesifo.de
Editor: Clemens Fuest
https://www.cesifo.org/en/wp
An electronic version of the paper may be downloaded
· from the SSRN website: www.SSRN.com
· from the RePEc website: www.RePEc.org
· from the CESifo website: https://www.cesifo.org/en/wp

CESifo *Working Paper No. 8890*

Adverse Effects of Control: Evidence from a Field Experiment

Abstract

We conduct a field experiment with remote workers to causally assess the impact of introducing a control mechanism on performance, and analyse the incidence of potential effects across tasks of different difficulty. We find that the implementation of control significantly reduces performance. The reduction occurs primarily among challenging tasks, whereas performance among simple tasks is unaffected. Further, the effects are primarily driven by workers with non-pecuniary motivation when uncontrolled. Our findings suggest that the relative importance of high performance among challenging tasks for employer profits can be a crucial determinant of the overall profitability of control interventions.

JEL Codes: C930, D210, J240, M500.

Keywords: control, hidden costs of control, remote work, field experiment.

*Holger Herz**
University of Fribourg
Fribourg / Switzerland
holger.herz@unifr.ch

Christian Zihlmann
University of Fribourg
Fribourg / Switzerland
christian.zihlmann@unifr.ch

*corresponding author

May 24, 2022
This RCT was registered before data collection on the AEA RCT registry (ID: AEARCTR-0003475 https://www.socialscienceregistry.org/trials/3475) and approved by the ethics committee of the Internal Review Board of the University of Fribourg (Ref-No.: 393). We are grateful for valuable comments to Björn Bartling, Berno Büchel, Alain Cohn, Martin Huber, Michael Kosfeld, Marina Schröder, Christian Zehnder, as well as seminar and conference participants in Kandersteg, Kiel, Dijon, Florida State University and the University of Fribourg. The authors declare that they have no other relevant or material financial interests that relate to the research described in this paper.

1 Introduction

The shift towards working from home has caused a huge rise in the demand for digital monitoring technology.[1] While such technology aims to increase performance of employees by reducing opportunities to slack (Alchian & Demsetz, 1972; Jensen & Meckling, 1976), the effectiveness of such measures has so far been under-explored, particularly concerning behavioral reactions to control that may potentially inhibit performance.[2] In May 2022, the Economist (2022) noted that "[...] many surveillance products aimed at boosting productivity are not well tested. Some risk being counterproductive."

To broaden our knowledge in this regard, we conduct a pre-registered natural field experiment with 693 remote workers on Amazon Mechanical Turk ("AMT"). AMT is an online crowdsourcing labor market where employers can recruit workers to perform short jobs for payment. The recruited workers for our job are unaware that they participate in an experiment and are tasked with extracting information from pictures showing game-play situations of a lacrosse match. The work process for each picture consists of two steps: First, workers need to declare whether or not the picture is readable, that is, whether they can extract information. Second, if the picture is declared as readable, workers need to extract information according to coding guidelines provided to them. If a picture is declared as unreadable, workers skip the second work step. Pictures vary in difficulty. While some are easy and the required information can be extracted quickly, others are more difficult and require substantial effort.

The experiment mimics the introduction of a novel control and surveillance mechanism in an existing work process and consists of two stages, a pre-treatment and an experimental stage. In the pre-treatment stage, workers receive a flat wage for working on 20 pictures. In the experimental stage, workers again face a set of 20 pictures and are randomly assigned to either the baseline or the treatment group.[3] Conditions in the baseline group ("Baseline") are identical to the pre-treatment stage. In the treatment group ("Controlled"), however, we communicate surveillance of the number of pictures that are declared as unreadable and implement a maximal allowance threshold: If workers declare

[1] Kropp (2021) reports that "During the pandemic, more than 1 out of 4 companies has purchased new technology, for the first time, to passively track and monitor their employees." Similar reports are found in Cutter, Chen, and Krouse (2020) and Hernandez (2020).

[2] Reports in newspapers suggest that increased workplace surveillance may increase stress and dissatisfaction of employees (see, for example, Blackman, 2020; Harwell, 2020).

[3] The set of 20 pictures used in the pre-treatment stage is different from the set in the experimental stage. In each stage, all workers are confronted with the same 20 pictures. For each stage, the order of appearance of the 20 pictures is randomly determined by the computer for each worker individually.

more than 8 out of the 20 pictures as unreadable, they do not receive the payment.[4]

Surveillance of the number of pictures declared as unreadable and the introduction of the maximum allowance threshold restrict workers' shirking possibilities by limiting the option to declare readable pictures as unreadable.[5] However, it only targets the first work step. A worker willing to shirk could simply declare pictures as readable, but then enter incomplete or random information in the second work step. Consequently, the control device is easily circumventable and thus expected to be ineffective in case workers want to act opportunistically.[6]

Yet, some workers may be motivated to perform in the employer's interest even if explicit performance incentives are weak and control is absent. Such non-pecuniary motivation could stem from, for example, gift-exchange (Akerlof & Yellen, 1990; Fehr, Kirchsteiger, & Riedl, 1993), an individual's desire to perform the task for its own sake (Bénabou & Tirole, 2003), a social norm (Sliwka, 2007), or pride and self-esteem (Ellingsen & Johannesson, 2008). If such non-pecuniary motivation is present among the workforce in absence of control, the implementation of control may have detrimental effects on performance (Frey, 1993; Falk & Kosfeld, 2006).

Our experimental setup allows us to advance our understanding of potential negative effects of the implementation of control along several dimensions: (1) We can causally assess potential adverse effects of control on worker performance in a remote work setting, (2) we can assess heterogeneity in reactions to control across workers with different degrees of non-pecuniary motivation when uncontrolled, and importantly (3) we can causally assess the incidence of potential negative effects across tasks of differing difficulty.

Our first result confirms that some workers reduce their performance when control is implemented, measured by correctly transcribed pictures. The average controlled worker reduces performance significantly by 5.5 percent relative to the Baseline ($p < .01$). We further find that control reduces the number of high performers. Whereas 40% of workers solve 14 pictures or more in the Baseline, only 30% of workers do so when controlled

[4]Note that this is a realistic feature in these type of tasks. Employers crowdsource such data entry tasks precisely because the requested information cannot be directly verified by the employer. Hence, controlling correct answers is not straightforward, whereas clicks on the unreadable button are easy to measure.

[5]We ensure that the minimum performance requirement is never a real constraint by including only two unreadable pictures in each set of 20 pictures.

[6]Note that such an ineffective control mechanism is representative of many types of control devices that are regularly observed in the field. Often, control can only be targeted on *observable* dimensions of the job, but shirking can simply be shifted into another, unobservable dimension. Examples include controlling the time logged into the employer network, but not productive working time, or controlling the number of calls made/received in a call center, but not the actual effort when on a call with a customer.

($p < .01$). At the same time, low performers are less common among controlled workers, but this effect is not significant at conventional levels. Jointly, these two effects imply that the variance of worker performance is significantly lower among controlled workers relative to the Baseline ($p < .05$). Put differently, control cultivates the average worker. We also find that control significantly reduces the time that workers are willing to invest to perform the job. Controlled workers invest on average 6.7% less time compared to the Baseline ($p < .05$).

Second, our data also allows us to study heterogeneous treatment effects within our population of workers. Workers are only randomized into Baseline and Controlled after completion of the pre-treatment stage. We measure workers pre-treatment motivation to perform the job by the time spent in the pre-treatment stage.[7] We find that the performance reduction in the Controlled treatment is particularly pronounced among workers that were motivated to perform in the pre-treatment stage, that is, those who invested relatively more time into solving the job. Splitting our sample at median pre-treatment motivation, we find that output among workers with high pre-treatment motivation is reduced by 1.1 pictures or 8.7% in Controlled, relative to the Baseline ($p < .01$). In contrast, workers with low pre-treatment motivation do not exhibit significant performance differences among the two groups. Two alternative proxies for non-pecuniary motivation — whether workers play lacrosse and whether workers re-consulted the coding guidelines while working in the pre-treatment stage — confirm these results. Thus, the implementation of control reduces the performance especially among those workers who were motivated to perform in the absence of control.[8]

Finally, we find that controlled workers reduce performance particularly among difficult and time-demanding tasks. Each worker transcribes the same set of 20 pictures, whereof 18 pictures are readable and not blurry. Ordering these readable pictures by the mean correct transcription rates in the Baseline and classifying them into three categories, we find that, compared to Baseline workers, controlled workers reduce performance by 20.4% among the hardest tertile ($p < .01$). We find a smaller worker performance reduction of 8.2% in the medium category ($p < .01$) and no significant difference for the easiest

[7] To identify non-pecuniary motivation, we rely as pre-registered on *procedural data* rather than outcome data such as performance, because procedural data is arguably more independent of worker's ability and other confounding factors that do not represent motivation. Refer to Section 2.3 for a more comprehensive elaboration.

[8] Note that this result does not imply that there is a crowding out of intrinsic motivation. Multiple potential behavioral mechanisms can explain why the implementation of control reduces performance especially among those workers who were motivated to perform in the absence of control. We elaborate further on this in the conclusion, see Section 4.

tertile. We find similar results when sorting the 20 pictures according to laboriousness, defined as the average time spent on a picture. Controlled workers perform significantly worse than Baseline workers among the more time-demanding pictures, reducing correct transcription rates by 11.7% ($p < .01$). Again, the decrease in worker performance is smaller among the medium (7.0% with $p < .05$) and the least labor-intensive pictures (3.1 % with $p < .10$). Control thus reduces worker performance among the hardest and most laborious tasks.

The finding that control differentially affects performance conditional on task difficulty has important implications. Ultimately, the adverse effects of control depend on the value a firm attaches to difficult tasks. For some firms, the value generated from solving a task may be uncorrelated or even negatively correlated with its difficulty. Take the example of a photographer who recruits crowdsource workers to identify bib numbers of runners to make images searchable by bib number on her/his website. If an image is slightly blurry, and thus hard to categorize, a runner, being the potential customer, is probably less likely to purchase it. It follows that difficult and time demanding images are arguably of low value to the employer. Consequently, performance is reduced at those tasks where it hurts the employer the least.

However, there are arguably many work environments in which the task difficulty positively correlates with the value generated for the employer. An example of such a situation is when the employer's aim is to train a machine learning algorithm, which is another common application of crowdsourced data entry tasks like the one used in our experiment.[9] Or more generally, tasks may be complementary inputs into production, as for example in the case of O-ring production functions (Kremer, 1993). In such instances, difficult to solve tasks (of which correct solutions are scarcest), have the highest marginal value to the firm. Thus, the fact that control reduces worker performance especially among those challenging tasks implies that the average treatment effect may strongly underestimate the impact of control on a firm's value of production.

Our findings shed light on the heterogeneity in the use of control mechanisms across different work environments (Ichniowski, Shaw, & Prennushi, 1997). In many jobs, workers have private information about the importance of different tasks for firm productivity, and firms cannot install control technology that accounts for this private information

[9]The literature documents that the introduction of falsely labelled training data can impede the accuracy of the deployed machine learning algorithm convexly. For instance, introducing 5% of mislabeled training data decreases the accuracy of the algorithm by 25% (Zhang & Yang, 2003). This is because the algorithm learns based on mislabeled data.

(Ichniowski & Shaw, 2003; Bartling, Fehr, & Schmidt, 2012). In such environments, one often observes high-performance work systems that refrain from control and instead grant authority to workers to prioritize tasks and solve problems themselves, without the necessity to follow strict processes. Our evidence shows that there can be good reason to refrain from controlling in such settings. Control reduces performance in particular among the difficult and laborious tasks. If task difficulty and the marginal value of a task are positively correlated, and both are private information of the worker, our data suggests that control could be highly detrimental for the firm.

Further, our findings contribute to the broader literature on adverse effects of control. Falk and Kosfeld (2006) were the first to show that control can undermine performance among certain agents. Their seminal article lead to a wave of subsequent laboratory work in this area, generally supporting the finding that control reduces performance of some agents, while the overall effect of control may be positive or negative (Kessler & Leider, 2016; Schnedler & Vadovic, 2011; Ziegelmeyer, Schmelz, & Ploner, 2012). For example, Masella, Meier, and Zahn (2014) show that adverse effects of control are present in between-group and withing-group matchings, and Riener and Wiederhold (2016) find that the adverse effects are more pronounced after a team building exercise.

The few studies on the effects of control that have been conducted in the field have been limited to tasks that do not differ in difficulty. When introducing control in unidimensional tasks, Nagin, Rebitzer, Sanders, and Taylor (2002) find that lowering the level of control leads most workers to decrease performance. Similarly, Boly (2011) finds that implementing control increases performance when tasks are unidimensional. Belot and Schröder (2016) investigate the effects of control in a framed field experiment and find that control increases performance in the monitored dimension, but decreases punctuality of workers, a non-monitored dimension.[10]

We go beyond these articles by studying heterogeneous treatment effects conditional on task difficulty, in addition to providing field evidence that control can have adverse effects in a remote work setting, an increasingly important work environment.[11] Because

[10]Similar to Belot and Schröder (2016), we find that shirking occurs in the non-controlled work step: The control device we implemented required workers to declare at least 12 out of the 20 pictures as "readable" in the first work step. Workers that intended to shirk would therefore sometimes have to declare pictures as readable, but could then enter random information into the entry form. Indeed, controlled workers are significantly less likely to declare pictures as unreadable, but the number of erroneous transcriptions significantly increases under control. Thus, workers reduce performance in the non-controlled work step.

[11]In 2015, only 44% of workers in the European Union conducted all their work at the employer's premises (Eurofound and the International Labour Office, 2017). During the Covid-19 outbreak from May to October 2020, half of all paid hours in the U.S. were provided from home. This trend towards

these environments are more impersonal, the observability of worker inputs is reduced and hence explicit control devices become more relevant. Intriguingly, Dickinson and Villeval (2008) and Schmelz and Ziegelmeyer (2020) find that performance reductions in response to the implementation of control are stronger when there is a more personal relationship between principal and agent. Thus, given that we study a remote work setting, our results probably constitute a lower bound in terms of the effect size.

The remainder of the paper proceeds as follows. Section 2 presents the experimental design and derives our hypotheses. Empirical tests and results are provided in Section 3. Finally, Section 4 concludes.

2 The Experiment

2.1 The real effort task

The natural field experiment is conducted on Amazon Mechanical Turk ("AMT"), an online crowdsourcing labor market where employers can recruit workers to perform short jobs for payment. Workers are not aware that they participate in a study. We appeared as a neutral AMT employer and did not reveal that we are researchers. Workers engage in a visual search task: extracting and categorizing information from a picture. Specifically, 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. Visual search tasks are common and natural on AMT and generate a productive output. Hence, workers sign up and engage in a job that fits their natural work environment. Figure 1 shows the entry form workers face.

For each picture, the first work step is to declare whether the picture is readable or not. Workers are instructed that a picture is defined as readable if it is not blurry and if all requested information is visible ("Clear image, all info visible"-button). Otherwise, the picture is not readable and workers need not to transcribe it ("Unclear image, not all info visible"-button).[12] If the picture is declared readable, workers have to enter five pieces of information in a second step. The entry form is shown in Figure 1. The job offers two ways in which workers can fail to correctly transcribe pictures, either by declaring

work from home is likely to persist (Barrero, Bloom, & Davis, 2020). In addition, the advent of the gig economy leads to a growing share of freelance work (De Stefano, 2016).

[12]Indeed, in some cases, declaring pictures as unreadable is the truthful response because the picture is blurry or some of the requested information is not identifiable, and workers knew that this may be the case. For this reason, such a button is a common feature in picture categorization tasks on AMT.

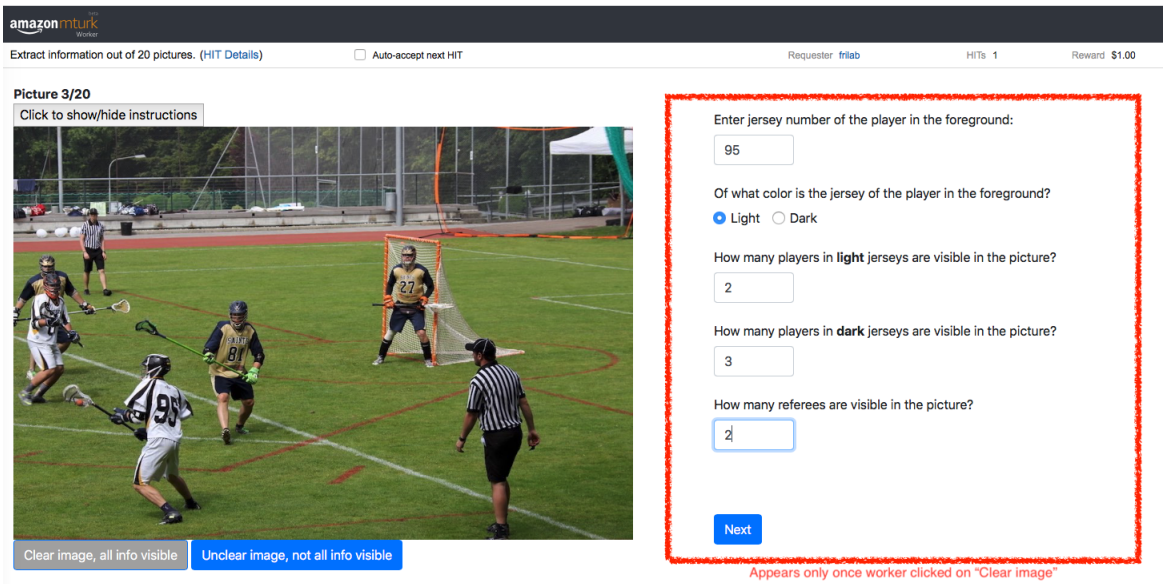


Figure 1: The real effort task

readable pictures as unreadable, or by correctly declaring readable pictures as readable but entering incomplete or incorrect information in the entry form.

An important feature of our design is that pictures vary in difficulty. While some pictures require little time to identify all relevant information and hence to transcribe them correctly, other pictures are cumbersome and require a substantial time investment (Figure B.1 in the Appendix provides examples of pictures of different difficulty).

2.2 Set-up and treatments

The experiment consists of two stages, a pre-treatment stage and an experimental stage.

2.2.1 The Pre-Treatment Stage

In the pre-treatment stage, all workers receive a flat payment of USD 1 for categorizing 20 pictures. Control is absent and any other form of extrinsic incentives is minimized: Workers are truthfully informed that the task is automatically approved and paid for regardless of the provided work ("All work is accepted: your job will be approved automatically within 1 day", which is an often used practice on AMT, see Appendix B for the full instructions). Consequently, workers can declare all 20 pictures as unreadable, not transcribing a single picture, and still receive the full reward.

The pre-treatment stage has a two-fold purpose in our experiment. First, it serves as a lock-in task with the goal to reduce dropouts once the treatment is induced. This is an established method on AMT to avoid selective attrition, see Horton, Rand, and

Zeckhauser (2011). Second, it allows us to observe behavior of all participants in the same environment with minimal extrinsic incentives.

2.2.2 The Experimental Stage

Once workers complete the pre-treatment stage, they are automatically offered the opportunity to do another set of 20 pictures. If workers accept the offer, they are randomized into one of two groups: The Baseline group receives the same contract as before, that is workers receive a flat payment of USD 1 for categorizing 20 pictures. The job is auto-approved and paid for regardless of the provided work. In contrast, the treatment group (henceforth: "Controlled") is assigned to a control mechanism: Workers are truthfully informed that they are allowed to declare a maximum of 8 out of 20 pictures as unclear and that this will be surveilled and verified automatically by the computer. If workers do not exceed the maximum allowance threshold of 8 pictures declared as unreadable, a flat reward of USD 1 is automatically paid. If the requirement is not met, workers are not eligible to receive the payment.[13]

The surveillance technology is intentionally bypassable: It targets only one way to shirk, namely declaring readable pictures as unreadable in the first work step. Shirkers can easily declare readable pictures as readable but enter erroneous information in the entry form in the second work step. This feature of our design allows us to observe the adverse effects of control, because disciplining effects are not induced.[14]

2.3 Measures, Procedures and Hypotheses

2.3.1 Measures

To produce a correct transcription of a picture, workers first need to identify readable pictures as "readable". Once done so, they also need to enter the correct information into the entry form. Hence, there are two ways in which a worker can fail to produce valuable output in our setting: (i) declaring a picture as unreadable even though it is readable, thus skipping it, or (ii) identifying a picture as readable, but entering erroneous information. To capture the first step of the work process, we define the variable SKIP

[13]The full instructions are available in Appendix B.

[14]Previous field studies such as Nagin et al. (2002) or Boly (2011) study the impact of control devices that have a disciplining effect. Thus their outcome measure is the net effect of the disciplining effect and any adverse effects of control. Because the disciplining effect is highly dependent on the specifics of the environment and the efficacy of the control device, a net positive effect of control does not imply the absence of adverse effects. Our setup allows us to isolate and quantify potential adverse effects.

as the number of pictures that are readable but declared as unreadable, and thus skipped by workers. To capture the second step of the work process, we define the variable ERRORS as the number of pictures that are declared as readable but the transcription is wrong. To capture overall work output, we define the variable OUTPUT as the total number of correctly solved pictures (note that there are 20 pictures in total and therefore: OUTPUT=20−SKIP−ERRORS).[15] OUTPUT thus represents worker performance and is our main variable of interest.

2.3.2 Procedures

The picture transcription task was programmed with the software oTree (Chen, Schonger, & Wickens, 2016). We conducted two randomized control trials, the first on December 10th 2018 and the second from March 9th to 11th 2020. Both trials were pre-registered before data collection (see Herz & Zihlmann, 2018). We conducted a second trial because we faced some missing data issues due to a software malfunction in the first trial, and because only a subset of our empirical analyses was pre-registered before the first trial.[16] In our analysis, we highlight those hypotheses for which adjustments in the pre-analysis plan were made between trial 1 and trial 2.

The total sample consists of 693 workers.[17] All workers were from the United States. We did not impose any other participation restriction. Workers received USD 1 for each stage. The mean duration to complete the job was about 7 minutes for each stage, yielding an hourly pay of approximately USD 9.

[15] In every set, two out of the 20 pictures are blurry and unreadable. Labeling the two unreadable pictures as unreadable is the truthful answer. Consequently, declaring an unreadable picture as unreadable is not contributing to SKIP nor to ERRORS but to OUTPUT.

[16] We focus the analysis on the pooled sample. All results remain qualitatively similar when analyzing the two trials separately. We report the separate analyses in Appendix C.

[17] The sample for the first trial consists of 203 workers and for the second trial it amounts to 490. 221 workers completed the first trial. We excluded 18 workers from the data set because they started the experimental stage more than once, thus being potentially familiar with both treatment conditions. There was no attrition after treatment induction: Every single worker who started the experimental stage also completed it. Note that workers learned about the treatment (i.e. that control is imposed, or not imposed) only once they started the experimental stage. Therefore, all workers continued with the job even though they were just informed that they are now subject to a control device. In the second trial, 512 workers completed the experimental stage. We excluded 22 workers from the data set either due to starting the experimental stage twice or because of failed attention checks that we included in the experimental procedure. We observed some attrition after treatment induction in the second trial. 43 workers learned about the treatment and started the experimental stage without completing it. Of those, 20 were assigned to the Baseline and 23 to the Controlled group. We thus deem attrition to be low and not significantly differently distributed across treatments. Dropped out Baseline and Controlled workers do not exhibit significant differences among any of the three performance dimensions.

2.3.3 Hypotheses

Our first hypothesis concerns the potential negative effect of implementing control on performance in our setting. The control technology used in the Controlled treatment restricts workers' shirking possibilities by limiting the option to declare pictures as unreadable, but it leaves the option open to erroneously and effortlessly transcribe the pictures. Hence, opportunistic agents can easily bypass the control technology and we do not expect a disciplining effect. On the other hand, if control is detrimental because workers react negatively to the implementation of control, controlled workers should reduce performance. Hypothesis 1 thus assesses the external validity of the laboratory finding that control entails hidden costs (Falk & Kosfeld, 2006).

Hypothesis 1. ***Control reduces performance.***[18] *Workers reduce performance when controlled.*

Our second hypothesis is concerned with heterogeneity across workers in their behavioral reaction to the control device. Frey (1993) posits that there are two types of agents, an opportunistic agent who always maximizes own income (or minimizes costs of effort), and an agent with non-pecuniary motivations who provides effort even in the absence of control or other types of extrinsic incentives.[19] Opportunistic agents should exert minimal effort and simply circumvent the control device. Those with non-pecuniary motivations, however, may react negatively to the implementation of control and reduce their effort (Carpenter & Myers, 2010; Dickinson & Villeval, 2008; Falk & Kosfeld, 2006). We therefore expect that the performance reduction when controlled is particularly pronounced among workers with non-pecuniary motivations.

Hypothesis 2. ***Control reduces performance among workers with non-pecuniary motivation.***[20] *The adverse effect of control is particularly pronounced among workers*

[18] Hypothesis 1 was pre-registered in both the analysis plans of study 1 and study 2 (Herz & Zihlmann, 2018).

[19] Non-pecuniary motivation could stem from, for example, reciprocity and gift-exchange (Akerlof & Yellen, 1990; Fehr et al., 1993), an individual's desire to perform the task for its own sake (Bénabou & Tirole, 2003), a social norm (Sliwka, 2007), or pride and self-esteem (Ellingsen & Johannesson, 2008).

[20] Hypothesis 2 was pre-registered in both the analysis plans of study 1 and study 2 (Herz & Zihlmann, 2018). The pre-analysis plans differ in the specification of the measurement of non-pecuniary motivation. In the pre-analysis plan for study 1, we pre-registered "playing or regularly watching lacrosse" as a proxy for non-pecuniary motivation for this job. However, few participants indicated that they play or regularly watch lacrosse resulting in limited power, and in-between the two pre-registrations an effective measurement for time spent on the task was developed for oTree. Hence, we adjusted our assessment and pre-registered for study 2 the time spent on the task in the pre-treatment stage as the proxy variable for non-pecuniary motivation. In Section 3, we provide results for both proxies, time on task and playing or regularly watching lacrosse.

with non-pecuniary motivation.

An important conceptual and empirical challenge in assessing this hypothesis is to ex ante identify those workers with higher non-pecuniary motivation. We adopt a broad and pragmatic concept of non-pecuniary motivation. The goal is to identify those workers who exert effort in absence of control. We thus consider workers to have high non-pecuniary motivation if they are motivated to act in the employer's interest in the pre-treatment stage when control is absent and explicit incentives weak.

We measure and employ labor input, that is the time devoted to our job in the pre-treatment stage when control is absent, as a proxy for non-pecuniary motivation.[21] Workers who devote more time to the job are classified as more motivated. We believe that time is a valid proxy for costly labor input because of the opportunity cost of time on AMT: Upon finishing, a worker can always switch to the next job and earn additional rewards. Thus, spending more time on our job is costly and reduces workers' hourly pay. Time represents *procedural data* and is thus arguably more independent of worker's experience, skills, cognitive ability and other confounding factors that do not represent motivation than work output measures such as performance.[22]

More precisely, we measure the time devoted to the task using `otree_tools` (Chapkovski & Zihlmann, 2019), which corrects for events in which workers switch away from the window in which the experiment is active and hence do not engage with the experimental job.[23]

We employ two alternative proxy variables to test the robustness of the results to hypothesis 2. First, we survey workers whether they play or regularly watch lacrosse. Workers who are familiar with the sport are assumed to be more motivated to do our job. Second, in the pre-treatment stage, we track whether workers re-consult the coding guidelines on how to classify pictures correctly while working on the picture classification job.[24] Workers who re-consult the guidelines are classified as workers with higher non-

[21] This follows the outlined approach in the pre-registration.

[22] See for example Carpenter and Huet-Vaughn (2019) for a discussion. Note also that time devoted to the task is correlated with performance (Spearman's $\rho = .09$, $p = .02$), as one would expect. Moreover, if performance measured through output is a noisy measure, employing pre-treatment output as a proxy for non-pecuniary motivation would result in a regression-to-the-mean problem. Indeed, when plotting a locally weighted regression of work output in the experimental stage against work output in the pre-treatment stage, we observe that initial low performers tend to better perform in stage 2. The opposite holds true for high performers. See Figure A.5 in the Appendix.

[23] Focus time has been shown to be a better predictor of work output than standard time (Chapkovski & Zihlmann, 2019).

[24] Workers could re-read the coding instructions by clicking on the "Click to show/hide instructions"-button, see Figure 1.

pecuniary motivation, because they strive to complete the task correctly according to the provided guidelines.

Our third hypothesis assesses heterogeneous reactions to control across types of tasks. Workers are tasked with transcribing 20 different pictures that vary considerably in their difficulty and in the amount of time required to solve them correctly. However, the control technology does not account for picture difficulty. This is why we hypothesize that the performance reduction should occur among those tasks at which effort costs are highest for the worker, and hence cost savings are highest when shirking. Consequently, we expect the control device to lead to a particularly pronounced performance reduction among challenging tasks.

Hypothesis 3. ***Control reduces performance among challenging tasks.***[25] *The adverse effect of control is particularly pronounced among the hard-to-solve pictures.*

3 Results

Table 1 provides descriptive statistics for our main outcome measures by treatment and stage. In the pre-treatment stage, Baseline workers solve on average 12.85 pictures correctly, and 12.03 in the experimental stage. Controlled workers on average solve 13.37 pictures correctly in the pre-treatment stage, and 11.87 in the experimental stage. Despite randomization into treatment, we thus observe a pre-treatment difference of 0.52 correctly solved pictures that is marginally significant at $p = .08$ (Welch's unequal variance t-test). Moreover, the general decrease in correctly solved pictures from the pre-treatment to the experimental stage is likely due to differences in the selection of pictures between the two stages, with the experimental stage being slightly more difficult. In the experimental stage, we observe the mean of skipped readable pictures (SKIP) to be 1.34 in Controlled, and only 3.8% of controlled workers skipped more than 8 pictures. Thus, the implemented control device was inconsequential for almost all workers regarding the eligibility to obtain the monetary reward.

As specified in the pre-registration, to account for potential pre-treatment differences, we control for individual, stage and time fixed effects in our subsequent analyses.[26] We do so by reporting our results as (i) the difference in our outcome variable between the

[25] This hypothesis was only pre-registered for the second trial, after exploratory findings in the first trial.

[26] This keeps individual factors such as ability, expertise, experience, fatigue and the device in use constant.

Table 1: Descriptive statistics

	Pre-treatment stage		Experimental stage		Difference	
	Baseline	Controlled	Baseline	Controlled	Baseline	Controlled
OUTPUT	12.85	13.37	12.03	11.87	-0.81	-1.50
	(4.05)	(3.67)	(3.94)	(3.54)	(2.73)	(2.47)
SKIP	2.51	2.14	2.00	1.34	-0.51	-0.79
	(2.86)	(2.44)	(3.01)	(2.20)	(2.18)	(1.54)
ERRORS	4.64	4.49	5.96	6.79	1.32	2.30
	(3.35)	(3.08)	(3.41)	(3.21)	(2.94)	(2.58)
Observations	693					

Note: The table displays the means along with the associated standard deviation (in parentheses) for the pre-treatment stage, the experimental stage, and the difference between the two stages. Note that workers were randomized into Baseline and Controlled only in the experimental stage. Thus, in the pre-treatment stage, workers were not yet assigned to a group. This implies that workers formed one group in the pre-treatment stage and were only randomly split into Baseline and Controlled in the experimental stage.

experimental and the pre-treatment stage (ii) as a regression approach by investigating the experimental stage outcomes conditional on the pre-treatment measurements.[27]

3.1 Control Reduces Worker Performance

In line with our pre-specified hypothesis, our first result establishes the existence of adverse effects of the implementation of control.

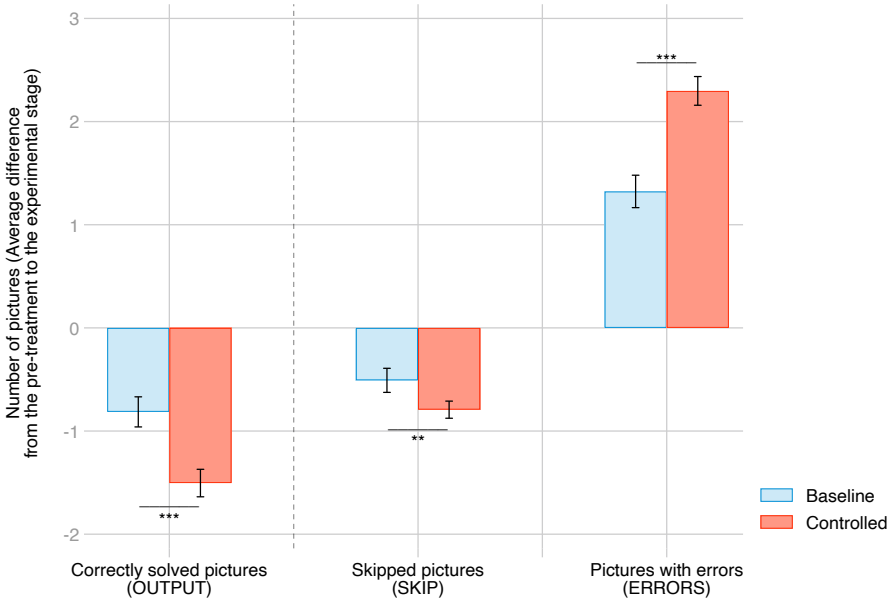
Result 1. *Control leads to a significant decrease in average work performance.*

Figure 2 provides support for Result 1. It shows that workers in the Baseline on average correctly solve 0.8 fewer pictures in the experimental stage than in the pre-treatment stage (variable OUTPUT). Notably, controlled workers decrease the number of correctly solved pictures by 1.5. This reduction is roughly twice as large as in the Baseline group and implies a significant difference of 0.7 additional unsolved pictures per worker relative to the Baseline ($p < .01$).[28] This is equivalent to a decrease of output of 5.5%. Thus, we find adverse effects of control in our sample.

[27] If treatment assignment is random, which it is in our case, both methods are unbiased (Breukelen, 2006; Wright, 2006) and reporting the results obtained from both methods is proposed to be a good practice (Allison, 1990; Lord, 1967).

[28] In this subsection, if not otherwise explicitly mentioned, when comparing two groups, we report p values from Welch's unpaired and two-sided t-test that accounts for unequal variances. When reporting p values from regressions, these are obtained from the OLS estimator employing robust standard errors.

Figure 2: Average treatment effect on workers' performance



Note: The graph reports on the vertical axis the number of pictures as an average difference from the pre-treatment to the experimental stage. Errors bars represent the standard error of the mean (accounting for unequal variances). OUTPUT: Number of correctly solved pictures. SKIP: Number of readable pictures that were declared as unreadable. ERRORS: Number of transcribed pictures that contain an error. $N = 693$, whereof Baseline $n = 350$, Controlled $n = 343$.
Welch's t-test p values: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

We test the robustness of our results by regressing experimental stage measurements on the treatment dummy while conditioning on the pre-treatment stage measurements to control for individual pre-treatment characteristics. Column (1) of Table 2 confirms Result 1. The control device reduces performance by 0.56 correctly solved pictures ($p < .01$).

Table 2: Regression Analysis: Average treatment effect on workers' performance

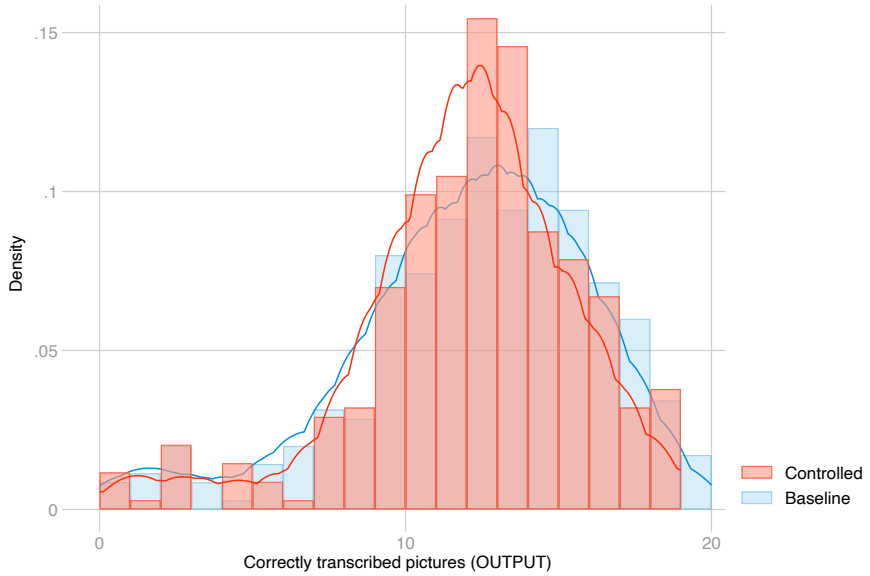
	(1) OUTPUT	(2) SKIP	(3) ERRORS
Controlled	-0.56	-0.38	0.92
	(0.18)	(0.13)	(0.19)
OUTPUT (pre-treatment)	0.74		
	(0.03)		
SKIP (pre-treatment)		0.74	
		(0.05)	
ERRORS (pre-treatment)			0.66
			(0.04)
Constant	2.49	0.14	2.89
	(0.42)	(0.13)	(0.23)
r2	0.59	0.56	0.42
N	693	693	693

Note: OLS regressions, robust standard errors (in parentheses). Outcome variables are experimental stage measurements. OUTPUT: Number of correctly solved pictures. SKIP: Number of readable pictures that were declared as unreadable. ERRORS: Number of transcribed pictures that contain an error. Pre-treatment variables of OUTPUT, SKIP and ERRORS control for the level of workers' performance before the treatment was induced.

We further find that control affects the distribution of performance in our workforce. Figure 3 depicts the distribution of correctly solved pictures for the Baseline and the Controlled treatment in the experimental stage. The kernel density estimates for controlled workers has more density around the mean of the distribution and flatter tails. Control therefore leads to both a lower frequency of low performing workers and a lower frequency of high performing workers. The distribution is significantly more centered around the mean, and Levene's test for the equality of variances reveals that, indeed, heterogeneity in worker performance is reduced by control ($p < .05$). Put differently, control cultivates the average worker.

Our data also allows us to identify the effect of control on time invested into the task, and we find that Controlled workers invest significantly less time into the job. Workers in the Baseline on average spend 6.9 minutes working on pictures in the pre-treatment stage and 7.25 minutes in the experimental stage. Consequently, those Baseline workers

Figure 3: Histogram and kernel density estimates of workers' performance



Note: The graph reports by experimental group a histogram of the variable OUTPUT (number of correctly transcribed pictures). The data are experimental stage measurements. The bin width is set to 1 because the data is discrete. Epanechnikov kernel density estimates are overlaid, the default (optimal) width was used.

on average work 20 seconds more in the experimental stage. Controlled workers invest on average 6.9 minutes in the pre-treatment stage, too, but only 6.75 minutes in the experimental stage. Controlled workers thus invest 9 seconds less in the experimental stage, which is a 30 seconds difference compared to Baseline workers who invest 20 seconds more. This yields a significant reduction of time invested by controlled workers of 7% compared to the Baseline ($p < .05$).

Figure 2 also provides insights about the reaction to control in the two steps of the work process, the number of skips and errors. Controlled workers reduce the number of skipped readable pictures by 0.8 between the pre-treatment stage and the experimental stage while non-controlled workers do so by 0.5 pictures only ($p < .05$). Simultaneously, we observe the number of transcribed pictures that contain errors to be 16.8% higher among controlled workers compared to the Baseline, a highly significant difference ($p < .01$).[29] Regression analysis (see columns (2) and (3) in Table 2) confirms these findings. Compared to the Baseline, controlled workers reduce the number of SKIPS on average by 0.38 pictures ($p < .01$) and increase the number of ERRORS on average by 0.92 ($p < .01$).

[29]This finding is robust to applying various alternative measurements for work quality, for example error rates instead of the absolute number of errors, errors by single input field instead binary by picture, and errors by single input field per picture (see Appendix Figure A.1). Controlled workers do not only transcribe more pictures erroneously, but also make more errors per picture.

Taken together, we find that the implementation of control decreases overall performance. The adverse effects of control arise in the non-controlled work step, whereas the performance metric in the controlled work step, the number of readable pictures that are declared unreadable, improves. This finding is related to Belot and Schröder (2016), who find that when workers are monitored in one dimension of a multidimensional effort task, performance in that dimension improves but decreases in other, unobserved dimensions. An implication of this result is that it can be difficult for firms to notice the detrimental impact of control, because control devices are necessarily implemented in observable dimensions or work steps, and performance metrics in those work steps only signal the positive effects of control.[30] However, the adverse effects of control may arise in other and potentially non-observed steps in the work process, in which performance metrics are absent.

3.2 Control Reduces Performance Among Workers with Non-pecuniary Motivation

Hypothesis 2 explores whether Result 1 is the consequence of a uniformly negative reaction to the control device or whether there is important heterogeneity in workers' behavioral response.

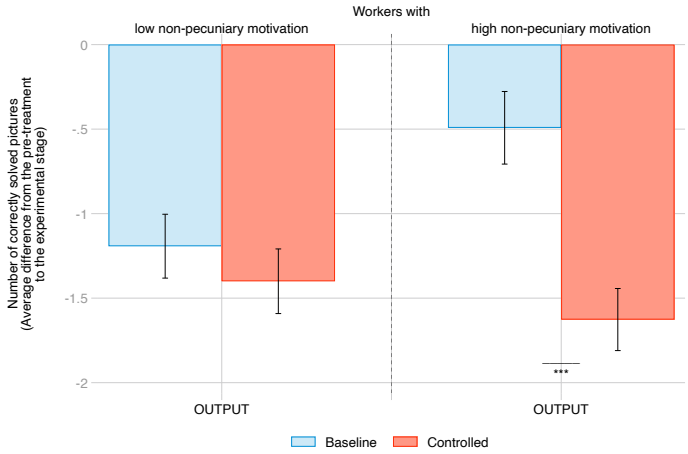
In particular, given that our control device could be easily circumvented, control should have no effect on the overall performance of opportunistic agents. Workers with non-pecuniary motivation, however, may reduce performance in reaction to control (Falk & Kosfeld, 2006; Ziegelmeyer et al., 2012). Our findings are summarized in Result 2.

Result 2. *The negative performance impact of control is significantly more pronounced among workers with high non-pecuniary motivation.*

Support for Result 2 can be seen in Figure 4. As explained in Section 2, we use pre-treatment labor input, captured by time spent on the job, as our measure of non-pecuniary motivation. We then classify workers into two types, those with high motivation and those with low motivation, based on a median split. Figure 4 plots the average difference of workers' performance between the pre-treatment stage and the experimental stage for both experimental groups and by both types of workers.

[30] In addition, because of the apparent positive feedback in the controlled step of the workflow, the performance metric may be misinterpreted and lead to false conclusions about the effectiveness of control.

Figure 4: Performance by type of worker



Note: The graph reports on the vertical axis the number of correctly solved pictures (OUTPUT) as an average difference from the pre-treatment to the experimental stage. Errors bars represent the standard error of the mean (accounting for unequal variances). Workers are classified into low and high non-pecuniary motivation based on a median split of pre-treatment work input (measured through time spent). Group sizes: Low non-pecuniary motivation $N = 346$, whereof Baseline $n = 161$, Controlled $n = 185$. High non-pecuniary motivation $N = 347$, whereof Baseline $n = 189$, Controlled $n = 158$.

Table 3: Regression Analysis: Non-pecuniary motivation interacted with treatment

	(1)	(2)
	OUTPUT	
Controlled	0.85	0.03
	(0.49)	(0.26)
Non-pecuniary motivation, cont	0.18	
	(0.05)	
Controlled × Non-pecuniary motivation, cont	-0.20	
	(0.07)	
Non-pecuniary motivation (=1)		0.98
		(0.27)
Controlled × Non-pecuniary motivation (=1)		-1.11
		(0.36)
OUTPUT (pre-treatment)	0.74	0.73
	(0.03)	(0.03)
Constant	1.33	2.06
	(0.45)	(0.41)
r2	0.60	0.60
N	693	693

Note: OLS regressions, robust standard errors (in parentheses). The outcome variable is the number of correctly solved pictures in the experimental stage (OUTPUT). Model (1) employs the continuous measurement of non-pecuniary motivation: Non-pecuniary motivation, cont is captured by work input in the pre-treatment stage which is measured through time spent (in minutes). Model (2) employs binary non-pecuniary motivation, resulting from a median split of work input: Workers are classified into low and high non-pecuniary motivation based on a median split of pre-treatment work input (measured through time). Pre-treatment OUTPUT controls for the level of workers' performance before the treatment was induced.

The right panel provides evidence supporting Result 2: Whereas motivated workers in the Baseline reduce their performance by approximately 0.5 pictures, motivated workers subject to a control device reduce output by 1.6 pictures, a highly significant difference of more than one picture. This is equivalent to a decrease of output by approx. 9% ($p < .01$) when motivated workers are controlled. For workers with low motivation, depicted in the left panel, we do not find significant differences in output between the two experimental groups. Moreover, the negative performance effect of control on motivated workers is significantly stronger than the negative performance effect of control on workers with low motivation ($p < .05$).

We continue with regression analysis to test the robustness of this result and regress our outcome variables of interest on individual non-pecuniary motivation. Column (1) in Table 3 measures non-pecuniary motivation continuously as the time spent on the task in the pre-treatment stage (in minutes). Note first that non-pecuniary motivation increases the number of correctly solved pictures among Baseline workers. Not so for controlled workers: the coefficient of the interaction term between the Controlled group dummy and non-pecuniary motivation is negative and statistically highly significant ($p < .01$). The higher the motivation of a worker, the stronger the negative reaction to control in our data. We observe the same pattern when median splitting workers into low and and high non-pecuniary motivation, see column (2). Again, the interaction term is negative and statistically highly significant ($p < .01$), indicating that workers with high non-pecuniary motivation are those that react especially adverse to the implementation of control.

Because non-pecuniary motivation is not exogenously varied, differences in the pre-treatment stage levels of motivation could be related to other factors. We thus test the robustness of Result 2 by employing two alternative proxies for non-pecuniary motivation, i) whether workers play or regularly watch lacrosse and ii) whether workers click the "Open Instructions"-button in the pre-treatment stage to re-consult the instructions on how to classify pictures properly. In the former case, 151 workers play or regularly watch lacrosse and are thus classified as motivated. Among those workers, performance is reduced by approx. 8.9% when controlled ($p < .10$) compared to the Baseline, while non-motivated workers do so by 4.7% only ($p < .01$).[31] In the latter case, 144 workers re-consulted the guidelines at least once, and are thus classified as motivated. Among those workers, performance is reduced by approx. 9.6% when controlled ($p < .01$) compared to the

[31]Note that for both alternative proxies, the group size of workers with low non-pecuniary motivation is substantially larger than the group size of workers with high non-pecuniary motivation. Statistical significance among the two types of workers is thus not directly comparable.

Baseline, while non-motivated workers do so by 4.2% only ($p < .05$). See Appendix A.2.2 for further detailed results.

Thus, for those two alternative proxies of non-pecuniary motivation, Result 2 holds: The performance reduction is particularly pronounced among motivated workers. Note that this finding does not imply that control crowds out intrinsic motivation. We will further elaborate on this in Section 4.

3.3 Control Reduces Worker Performance Among Challenging Tasks

We now turn to our third hypothesis, positing that the performance reduction particularly arises in more challenging tasks. To assess this hypothesis, we take advantage of the experimental design that tasks workers with transcribing pictures of varying difficulty.[32] We proceed as detailed in the pre-registration and categorize the 18 pictures that were readable into three categories based on their difficulty, measured by the achieved performance (OUTPUT).[33] The categorization is based on the performance of the Baseline group. Our findings are summarized in Result 3.

Result 3. *The negative performance impact of control is significantly more pronounced among hard-to-solve pictures.*

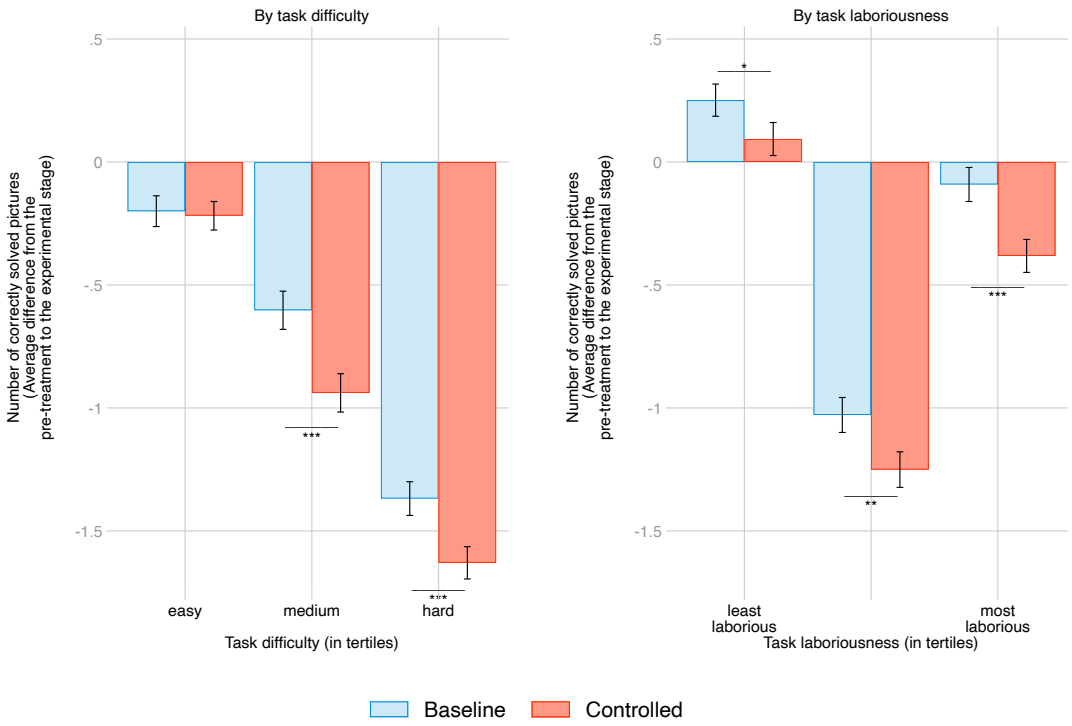
Support for Result 3 is shown in Figure 5, which plots the average difference of correctly solved pictures by picture difficulty and experimental group. In the left panel, the leftmost bars show that the control device hardly affects correct transcriptions of easy-to-solve pictures. In the medium category however, Baseline workers solve 0.6 fewer pictures in the experimental stage than in the pre-treatment stage, while controlled workers solve 0.9 fewer pictures. Controlled workers thus perform worse than the Baseline by 0.3 pictures or 8.2% ($p < .01$). Among hard pictures, this treatment effect grows in magnitude. Controlled workers perform worse compared to the Baseline by 0.26 pictures, which represents a substantial performance reduction of 20.4% ($p < .01$).

The right panel in Figure 5 plots a similar graph but by task laboriousness instead of task difficulty: Pictures are ordered into laboriousness tertiles based on the average time spent on a picture in the Baseline group. Interestingly, a very similar pattern emerges.

[32] Note that order of pictures is randomly determined by the computer for each worker individually.

[33] As pre-registered, we exclude the two blurry and unreadable pictures for the analysis because as expected, these two pictures are correctly classified as unreadable by the vast majority of the workforce. Excluding these two pictures allows us to create three categories that represent difficulty tertiles.

Figure 5: Performance by task heterogeneity



Note: The graph reports on the vertical axis the number of correctly transcribed pictures (OUTPUT) as an average difference from the pre-treatment to the experimental stage, representing the change in performance. The left panel reports the performance difference by task difficulty, the lower panel by task laboriousness. For each stage separately, pictures are classified into difficulty tertiles based on the performance of the Baseline group and into task laboriousness tertiles based on the time elapsed of the Baseline group. $N = 693$, whereof Baseline $n = 350$, Controlled $n = 343$.

We observe that the performance reduction of controlled workers is especially pronounced among pictures that require more labor. While the performance reduction of controlled workers compared to non-controlled workers amounts to 0.15 pictures or 3.1% in the least laborious category ($p < .10$), it amounts to 0.22 pictures or 7.0% in the medium category ($p < .05$) and to 0.29 pictures or 11.7% among the most labor-intensive pictures ($p < .01$).

To assess the robustness of our results, we turn to regression analysis and estimate the models shown in Table 4. Column (1) to (3) report the regression coefficients when pictures are classified into three categories based on their difficulty. In the easy picture category (1), controlled workers do not perform worse than Baseline workers. The adverse effects of control occur among the medium (column (2)) and hard pictures (column (3)). The control device reduces performance in the medium picture category by 0.25 pictures ($p < .05$) and in the hard picture category by 0.24 pictures ($p < .01$), conditional on the pre-treatment performance. Again, similar results emerge when we order pictures according to task laboriousness. Workers do not differ among the least time-demanding pictures, but controlled workers reduce performance by 0.17 pictures among the medium laborious category ($p < .10$) and substantially by 0.25 pictures among the labor-intensive tasks ($p < .01$).

Table 4: Regression Analysis: Performance by task heterogeneity

	(1)	(2)	(3)	(4)	(5)	(6)
	OUTPUT					
	by task difficulty			by task laboriousness		
	easy	medium	hard	least	medium	most
Controlled	-0.01	-0.25	-0.24	-0.05	-0.17	-0.25
	(0.08)	(0.10)	(0.08)	(0.08)	(0.09)	(0.09)
OUTPUT (pre-treatment)	0.88	0.67	0.59	0.62	0.56	0.63
	(0.07)	(0.03)	(0.03)	(0.04)	(0.03)	(0.03)
Constant	0.41	0.84	-0.31	2.03	0.71	0.74
	(0.38)	(0.16)	(0.08)	(0.23)	(0.14)	(0.08)
r2	0.39	0.37	0.37	0.42	0.32	0.40
N	693	693	693	693	693	693

Note: OLS regressions, robust standard errors (in parentheses). Outcome variables are the experimental stage measurements of the number of correctly solved pictures (OUTPUT) by task difficulty and by task laboriousness, respectively. The 18 readable pictures are classified into three categories by task difficulty based on the number of correctly solved pictures and into three categories by task laboriousness based on the time spent on a picture. The specification controls for the level of workers' pre-treatment performance (OUTPUT) in the respective category.

We also find that the performance reduction among hard and labor-intensive tasks is primarily driven by the motivated workforce. Figure A.6 in the Appendix provides

support. Compared to the Baseline, controlled workers with low non-pecuniary motivation actually perform slightly better in the easy picture category ($p < .10$), not differently in the medium picture category, and slightly worse among hard-to-solve pictures ($p < .10$).

In contrast, controlled workers with high non-pecuniary motivation significantly reduce performance in all pictures categories. The magnitude of the effect amounts to 0.25 pictures or 4.7% among easy pictures ($p < .05$), to 0.57 pictures or 13.3% among medium pictures ($p < .01$) and to 0.27 pictures or 31.5% among challenging pictures ($p < .05$).

To sum up, we provide evidence that the implementation of control reduces performance particularly among the hard-to-solve and labor-intensive tasks. This result not only introduces a novel facet regarding the adverse reactions to control, but also has important implications for firms. Given that performance reductions primarily arise in challenging tasks, the effect of control on a firm's profitability will crucially depend on the importance of those tasks for the firm. In particular in work environments in which different tasks are complements, difficult tasks are likely of highest marginal value for the firm. Thus, the average treatment effect on performance may substantially underestimate the impact of control on firm profitability.

4 Conclusion

This article provides novel evidence on the adverse effects of control from a field experiment conducted in a work from home setting. We document that control adversely affects worker performance, in particular among difficult tasks and among workers with non-pecuniary motivation when uncontrolled. Our findings imply that the implementation of control can be profoundly harmful (1) for firms whose workforce is motivated to perform even if extrinsic incentives are mainly absent, and (2) for firms that receive particularly high marginal value from worker performance among challenging tasks.

The latter finding aligns with the observation that control mechanisms are rarely used in high-performance work-systems (Ichniowski & Shaw, 1999). In high-performance work-systems, employers rely on worker's private information to identify those tasks that are particularly valuable to the firm. Because this information is private, the employer cannot implement control mechanisms that account for it. Our results suggest that implementing imperfect control devices can be particularly detrimental in such instances because, under the plausible assumption that marginal value and difficulty of the task are positively correlated, control causes performance reductions at precisely those highly valuable tasks.

At the same time, our findings do not imply that control is always detrimental. We deliberately implemented a control device that workers could easily circumvent, because the focus of this paper was on identifying the adverse effects of control. Our results show that moderately effective control devices will likely have positive overall performance effects, in particular when tasks are perfect substitutes.

Our observed treatment effect is potentially consistent with multiple behavioral theories, in particular reciprocity (Akerlof & Yellen, 1990; Fehr et al., 1993), intrinsic motivation (Bénabou & Tirole, 2003), conformity with social norms (Sliwka, 2007), or pride and self-esteem (Ellingsen & Johannesson, 2008). While our experimental design does not allow us to cleanly discriminate among these potential mechanisms, our data still provides some insights. In particular, crowding out of intrinsic motivation is a reduction of the enjoyment one derives from performing the activity itself (Deci, 1971; Bénabou & Tirole, 2003) and appears to be an unlikely explanation of our findings. At the end of the experimental stage, we asked workers "What reward would be appropriate for doing your work?" in order to elicit workers demand of renumeration for performing the job. If the task was intrinsically less rewarding in the Controlled group (and thus intrinsic motivation was crowded out), workers derived less utility from performing the task itself, and as a consequence, workers should indicate a higher renumeration as appropriate. Yet, we do not find such an effect. The median requested monetary reward for doing the job is exactly the same for both groups, namely USD 1.75.[34] This aligns with laboratory evidence rejecting crowding out of intrinsic motivation as a main driver of the adverse effects of control (Dickinson & Villeval, 2008).

Moreover, the behavioral heterogeneity in our data has important implications for the design of organizations. Ultimately, how can an organization design incentives schemes that discipline the opportunistic workers without reducing performance of those with non-pecuniary motivation? In this respect, it is important to note that our findings relate to a situation in which control is newly and uniformly implemented within the existent workforce of a firm. Such an implementation can be interpreted as a signal of distrust from the employer, which may be one potential mechanism that causes the adverse effects of control (see, for example, Sliwka, 2007; Ellingsen & Johannesson, 2008). The results do not necessarily generalize to situations in which workers start working in a firm that either already uses control technology, or to situations in which only a part of the workforce is

[34] Median test: $p = .93$, Mann-Whitney-U test: $p = .76$, Welch's t-test: $p = .19$. The result also holds when controlling for the time spent on the task.

confronted with a control device. Such settings would be interesting to study in further research.

More generally, the existence of different control regimes across and within firms raises interesting additional questions in terms of behavioral reactions of workers but also in terms of worker selection. Kosfeld and Von Siemens (2011) show that separating equilibria can exist for opportunistic and conditionally cooperative workers. Do separating equilibria also exist for intrinsically and extrinsically motivated workers? The literature documents self-selection with regard to other behavioral factors, such as overconfidence (Larkin & Leider, 2012), or a preference for being one's own boss (Hamilton, 2000; Hurst & Pugsley, 2011; Bartling, Fehr, & Herz, 2014). It is an interesting empirical question whether some workers would, for example, be willing to forgo monetary compensation in exchange for less control and more autonomy.

References

Akerlof, G. A., & Yellen, J. L. (1990, 05). The Fair Wage-Effort Hypothesis and Unemployment*. *The Quarterly Journal of Economics*, *105*(2), 255-283.

Alchian, A. A., & Demsetz, H. (1972). Production, information costs, and economic organization. *The American Economic Review*, *62*(5), 777–795.

Allison, P. D. (1990). Change scores as dependent variables in regression analysis. *Sociological Methodology*, *20*, 93–114.

Barrero, J. M., Bloom, N., & Davis, S. J. (2020, December). *Why working from home will stick* (Working Paper NO. 2020-174). University of Chicago.

Bartling, B., Fehr, E., & Herz, H. (2014). The intrinsic value of decision rights. *Econometrica*, *82*(6), 2005–2039.

Bartling, B., Fehr, E., & Schmidt, K. M. (2012, April). Screening, competition, and job design: Economic origins of good jobs. *American Economic Review*, *102*(2), 834-64.

Belot, M., & Schröder, M. (2016). The spillover effects of monitoring: A field experiment. *Management Science*, *62*(1), 37-45.

Blackman, R. (2020, May). How to monitor your employees — while respecting their privacy. *Harvard Business Review*. (https://hbr.org/2020/05/how-to-monitor-your-employees-while-respecting-their-privacy)

Boly, A. (2011, May 01). On the incentive effects of monitoring: evidence from the lab and the field. *Experimental Economics*, *14*(2), 241–253.

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.

Bénabou, R., & Tirole, J. (2003, 07). Intrinsic and Extrinsic Motivation. *The Review of Economic Studies*, *70*(3), 489-520.

Carpenter, J., & Huet-Vaughn, E. (2019). Real-effort tasks. In *Handbook of research methods and applications in experimental economics.* Cheltenham, UK: Edward Elgar Publishing.

Carpenter, J., & Myers, C. K. (2010). Why volunteer? evidence on the role of altruism, image, and incentives. *Journal of Public Economics*, *94*(11), 911-920.

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.

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.

Cutter, C., Chen, T.-P., & Krouse, S. (2020, Apr). *You're Working From Home, but Your Company Is Still Watching You*. (https://www.wsj.com/articles/youre-working-from-home-but-your-company-is-still-watching-you-11587202201)

Deci, E. L. (1971). Effects of externally mediated rewards on intrinsic motivation. *Journal of personality and Social Psychology*, *18*(1), 105.

De Stefano, V. (2016). *The rise of the "just-in-time workforce": On-demand work, crowdwork and labour protection in the "gig- economy"* (Tech. Rep.).

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.

Economist. (2022, May). Welcome to the era of the hyper-surveilled office. *The Economist*. (https://www.economist.com/business/welcome-to-the-era-of-the-hyper-surveilled-office/21809219)

Ellingsen, T., & Johannesson, M. (2008, 06). Pride and prejudice: The human side of incentive theory. *The American Economic Review*, *98*(3), 990-1008.

Eurofound and the International Labour Office. (2017). *Working anytime, anywhere: The effects on the world of work* (Tech. Rep.).

Falk, A., & Kosfeld, M. (2006). The hidden costs of control. *The American Economic Review*, *96*(5), 1611-1630.

Fehr, E., Kirchsteiger, G., & Riedl, A. (1993, 05). Does Fairness Prevent Market Clearing? An Experimental Investigation*. *The Quarterly Journal of Economics*, *108*(2), 437-459.

Frey, B. S. (1993). Does monitoring increase work effort? The rivalry with trust and loyalty. *Economic Inquiry*, *31*(4), 663–670.

Hamilton, B. H. (2000). Does entrepreneurship pay? an empirical analysis of the returns to self-employment. *Journal of Political economy*, *108*(3), 604–631.

Harwell, D. (2020, April). Managers turn to surveillance software, always-on webcams to ensure employees are (really) working from home. *The Washington Post*. (https://www.washingtonpost.com/technology/2020/04/30/work-from-

home-surveillance)

Hernandez, K. (2020, Mar). Even if you're working from home, your employer is still keeping track of your productivity—here's what you need to know. *CNBC*. (https://www.cnbc.com/2020/03/19/when-working-from-home-employers-are-watching—heres-what-to-know.html)

Herz, H., & Zihlmann, C. (2018). Does monitoring adversely affect worker performance? evidence from a natural field experiment. *AEA RCT Registry. October 23*.

Horton, J. J., Rand, D. G., & Zeckhauser, R. J. (2011, Sep 01). The online laboratory: conducting experiments in a real labor market. *Experimental Economics*, *14*(3), 399–425.

Hurst, E., & Pugsley, B. W. (2011). What do small businesses do? *Brookings Papers on Economic Activity*, 73–143.

Ichniowski, C., & Shaw, K. (1999). The effects of human resource management systems on economic performance: An international comparison of u.s. and japanese plants. *Management Science*, *45*(5), 704-721.

Ichniowski, C., & Shaw, K. (2003, March). Beyond incentive pay: Insiders' estimates of the value of complementary human resource management practices. *Journal of Economic Perspectives*, *17*(1), 155-180.

Ichniowski, C., Shaw, K., & Prennushi, G. (1997). The effects of human resource management practices on productivity: A study of steel finishing lines. *The American Economic Review*, *87*(3), 291–313.

Jensen, M. C., & Meckling, W. H. (1976). Theory of the firm: Managerial behavior, agency costs and ownership structure. *Journal of financial economics*, *3*(4), 305–360.

Kessler, J., & Leider, S. (2016). Procedural fairness and the cost of control. *The Journal of Law, Economics, and Organization*, *32*(4), 685-718.

Kosfeld, M., & Von Siemens, F. A. (2011). Competition, cooperation, and corporate culture. *The RAND Journal of Economics*, *42*(1), 23–43.

Kremer, M. (1993). The o-ring theory of economic development. *The Quarterly Journal of Economics*, *108*(3), 551–575.

Kropp, B. (2021, January). 9 trends that will shape work in 2021 and beyond. *Harvard Business Review*. (https://hbr.org/2021/01/9-trends-that-will-shape-work-in-2021-and-beyond)

Larkin, I., & Leider, S. (2012, May). Incentive schemes, sorting, and behavioral biases of

employees: Experimental evidence. *American Economic Journal: Microeconomics*, *4*(2), 184-214.

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

Masella, P., Meier, S., & Zahn, P. (2014). Incentives and group identity. *Games and Economic Behavior*, *86*(Supplement C), 12 - 25.

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.

Riener, G., & Wiederhold, S. (2016). Team building and hidden costs of control. *Journal of Economic Behavior & Organization*, *123*(Supplement C), 1 - 18.

Schmelz, K., & Ziegelmeyer, A. (2020). Reactions to (the absence of) control and workplace arrangements: experimental evidence from the internet and the laboratory. *Experimental Economics*, *23*(4), 933–960.

Schnedler, W., & Vadovic, R. (2011). Legitimacy of control. *Journal of Economics & Management Strategy*, *20*(4), 985–1009.

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.

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.

Zhang, J., & Yang, Y. (2003). Robustness of regularized linear classification methods in text categorization. In *Proceedings of the 26th annual international acm sigir conference on research and development in informaion retrieval* (pp. 190–197).

Ziegelmeyer, A., Schmelz, K., & Ploner, M. (2012, Jun 01). Hidden costs of control: four repetitions and an extension. *Experimental Economics*, *15*(2), 323–340.

Appendices

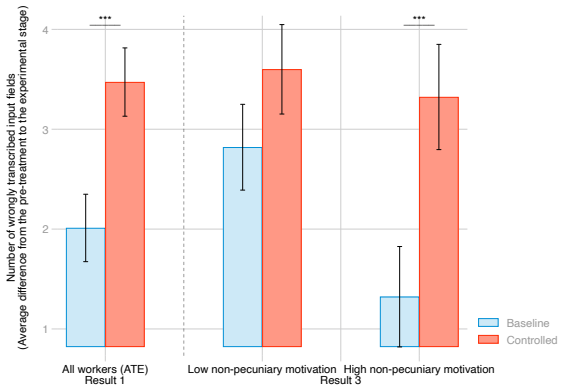
A Further Results

A.1 Control Reduces Performance

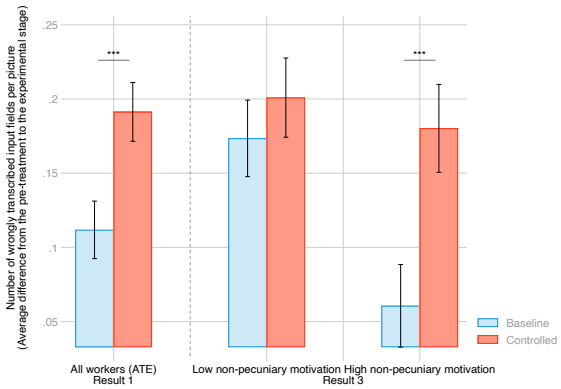
Figure A.1 employs alternative measures for ERRORS. Instead of a dichotomous classification of a picture as correct or false, Figure A.1a reports the average number of wrongly transcribed input fields (there are five input fields per picture), Figure A.1b reports the number of wrongly transcribed input fields per attempted picture (that is per non-skipped picture) and Figure A.1c reports the number of pictures that contain an error divided by the total of attempted, non-skipped pictures, thus representing the number of attempted pictures that contain at least one error.

Figure A.1: Alternative measures for ERRORS

(a) Number of wrongly transcribed input fields



(b) Number of wrongly transcribed input fields per attempted picture



(c) Percentage of pictures with errors

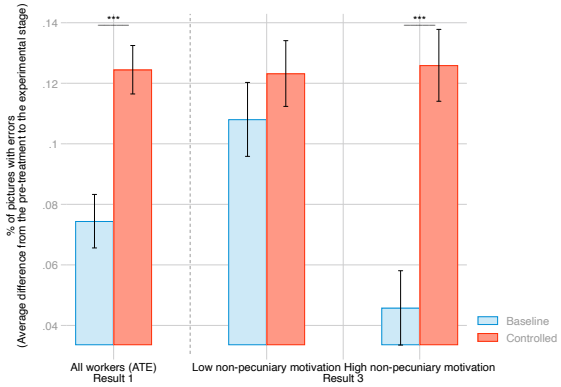


Figure A.2: Performance by type of worker, proxied by the time spent in the pre-treatment stage

Note: The graph reports on the vertical axis the number of pictures as an average difference from the pre-treatment to the experimental stage. Errors bars represent the standard error of the mean (accounting for unequal variances). OUTPUT: Number of correctly solved pictures. SKIP: Number of readable pictures that were declared as unreadable. ERRORS: Number of transcribed pictures that contain an error. Workers are classified into low and high non-pecuniary motivation based on a median split of pre-treatment work input (measured through time spent on task). Group sizes: Low non-pecuniary motivation N=346, whereof Baseline n=161, Controlled n=185. High non-pecuniary motivation N=347, whereof Baseline n=189, Controlled n=158.
Welch's t-test p values: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

A.2 Control Reduces Performance Among Workers With Non-pecuniary Motivation

A.2.1 Performance by type of worker

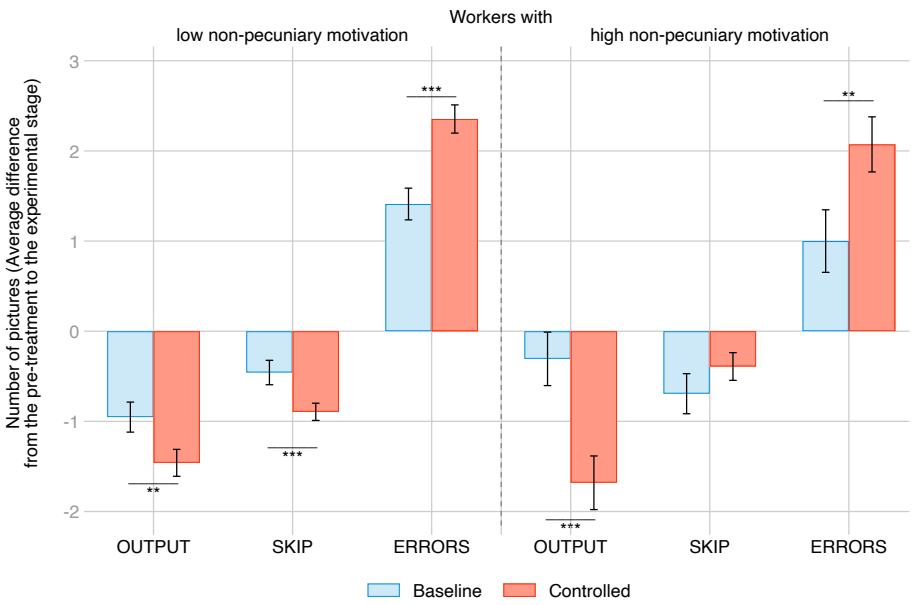
A.2.2 Alternative Proxy Variables for Non-Pecuniary Motivation

We test the robustness of Result 2 with alternative proxy variables for non-pecuniary motivation. Figure A.3 shows results when workers are classified into two types, those with high motivation and those with low motivation, based on whether they re-consulted in the pre-treatment stage the instructional guidelines of the picture transcription job. Workers who re-consulted the instructions are classified as those with higher non-pecuniary motivation. Note that this is not a median split and the group of workers with low motivation is substantially larger. Hence, statistical significance is harder to compare among the

two types of workers. Figure A.3 plots the average differences in our outcome variables between the pre-treatment stage and the experimental stage for both experimental groups and by both types of workers.

The leftmost bars in the right panel display the number of correctly solved pictures and provides evidence supporting Result 2: Whereas motivated workers in the Baseline reduce their output by approximately 0.3 pictures, motivated workers subject to a control device reduce output by 1.7 pictures, a highly significant difference of more than one picture, equivalent to a performance decrease by approx. 9.6% ($p < .01$). For workers with low motivation, depicted in the left panel, the performance decrease only amounts to approx. 4.2% ($p < .05$).

Figure A.3: Performance by type of worker, proxied by click on "Open Instructions"-button



Note: The graph reports on the vertical axis the number of pictures as an average difference from the pre-treatment to the experimental stage. Errors bars represent the standard error of the mean (accounting for unequal variances). The horizontal axis plots work output, representing workers' performance, and its two dimensions. OUTPUT: Number of correctly solved pictures. SKIP: Number of readable pictures that were declared as unreadable. ERRORS: Number of transcribed pictures that contain an error. Workers are considered as workers with high non-pecuniary motivation when they re-consulted the classification instructions at least once in the pre-treatment stage. All other workers are considered to be of low non-pecuniary motivation. Group sizes: Low non-pecuniary motivation N=549, whereof Baseline n=275, Controlled n=274. High non-pecuniary motivation N=144, whereof Baseline n=75, Controlled n=69.
Welch's t-test p values: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Regression table A.1 confirms this result. The interaction term of the Controlled

Table A.1: Regression Analysis: Non-pecuniary motivation proxied by clicks on the "Open Instructions" -button

	(1) OUTPUT	(2) SKIP	(3) ERRORS
Controlled	-0.39	-0.54	0.93
	(0.21)	(0.16)	(0.22)
Non-pecuniary motivation (=1)	0.92	-0.43	-0.50
	(0.32)	(0.23)	(0.35)
Controlled × Non-pecuniary motivation (=1)	-0.74	0.77	-0.08
	(0.44)	(0.27)	(0.48)
OUTPUT (pre-treatment)	0.74		
	(0.03)		
SKIP (pre-treatment)		0.74	
		(0.05)	
ERRORS (pre-treatment)			0.66
			(0.04)
Constant	2.38	0.24	3.02
	(0.43)	(0.15)	(0.23)
r2	0.59	0.57	0.43
N	693	693	693

Note: OLS regressions, robust standard errors (in parentheses). Outcome variables are experimental stage measurements. OUTPUT: Number of correctly solved pictures. SKIP: Number of readable pictures that were declared as unreadable. ERRORS: Number of transcribed pictures that contain an error. A worker is classified as non-pecuniary motivated if he or she clicked at least once the ‘ Open Instructions ’-button in the pre-treatment stage, allowing the worker to reconsult the picture classification instructions. Pre-treatment variables of OUTPUT, SKIP and ERRORS control for the level of workers’ performance before the treatment was induced.

treatment and non-pecuniary is negative, meaning that Controlled workers that were classified as non-pecuniary motivated decrease OUTPUT more than others ($p < .10$).

Figure A.4 shows results when workers are classified into two types, those with high motivation and those with low motivation, based on whether they either play or regularly watch lacrosse (or both). Workers familiar with the sport are assumed to have higher non-pecuniary motivation. Workers that do not play or regularly watch lacrosse are classified as workers with low non-pecuniary motivation. Note that again, this is not a median split and the group of workers with low motivation is substantially larger. Figure A.4 plots the average differences in our outcome variables between the pre-treatment stage and the experimental stage for both experimental groups and by both types of workers.

Again, we find evidence supporting supporting Result 2: Whereas motivated workers in the Baseline increase their output by approximately 0.1 pictures, motivated workers subject to a control device reduce output by 0.85 pictures, a significant difference equivalent to a performance decrease when controlled by approx. 8.9% ($p < .10$). For workers with low motivation, depicted in the left panel, the performance decrease under control only amounts to approx. 4.7% ($p < .01$).

Table A.2: Regression Analysis: Non-pecuniary motivation proxied by familiarity with the sport lacrosse

	(1) OUTPUT	(2) SKIP	(3) ERRORS
Controlled	-0.53	-0.39	0.92
	(0.20)	(0.14)	(0.21)
Non-pecuniary motivation (=1)	0.22	0.18	-0.17
	(0.37)	(0.33)	(0.38)
Controlled × Non-pecuniary motivation (=1)	-0.12	0.02	0.03
	(0.51)	(0.39)	(0.53)
OUTPUT (pre-treatment)	0.75		
	(0.03)		
SKIP (pre-treatment)		0.73	
		(0.05)	
ERRORS (pre-treatment)			0.67
			(0.04)
Constant	2.35	0.12	2.90
	(0.44)	(0.13)	(0.23)
r2	0.59	0.56	0.42
N	693	693	693

Note: OLS regressions, robust standard errors (in parentheses). Outcome variables are experimental stage measurements. OUTPUT: Number of correctly solved pictures. SKIP: Number of readable pictures that were declared as unreadable. ERRORS: Number of transcribed pictures that contain an error. A worker is classified as non-pecuniary motivated if he or she plays or regularly watches lacrosse. Pre-treatment variables of OUTPUT, SKIP and ERRORS control for the level of workers' performance before the treatment was induced.

Figure A.4: Performance by type of worker, proxied by familiarity with the sport lacrosse

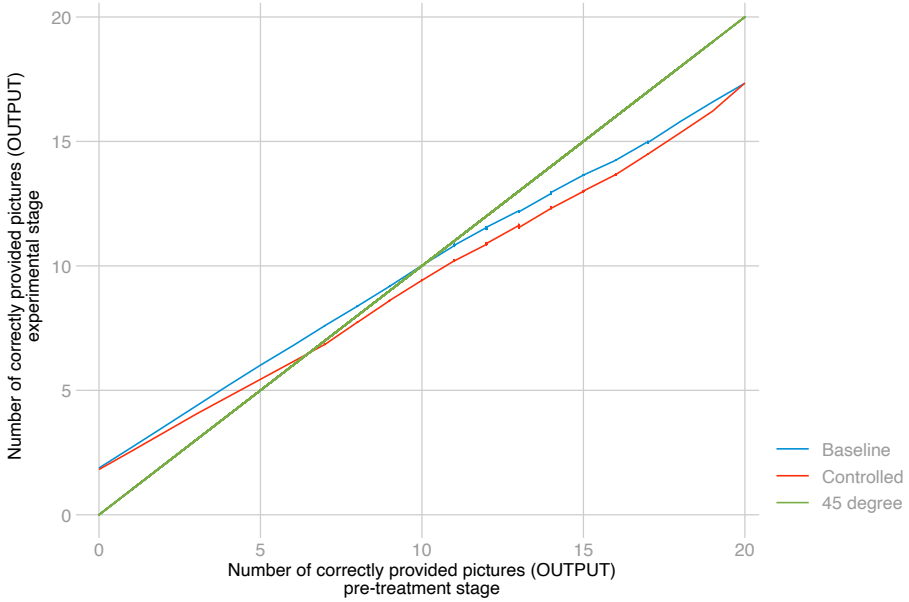
Note: The graph reports on the vertical axis the number of pictures as an average difference from the pre-treatment to the experimental stage. Errors bars represent the standard error of the mean (accounting for unequal variances). The horizontal axis plots work output, representing workers' performance, and its two dimensions. OUTPUT: Number of correctly solved pictures. SKIP: Number of readable pictures that were declared as unreadable. ERRORS: Number of transcribed pictures that contain an error. Workers are classified into high non-pecuniary motivation if workers either play or regularly watch lacrosse (or both). All other workers who are unfamiliar with the sport are classified into low non-pecuniary motivation. Group sizes: Low non-pecuniary motivation N=542, whereof Baseline n=274, Controlled n=268. High non-pecuniary motivation N=151, whereof Baseline n=76, Controlled n=75.
Welch's t-test p values: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

In regression table A.2, we observe the interaction term of the Controlled treatment and non-pecuniary motivation to be negative. Again, this means that Controlled workers that were classified as non-pecuniary motivated because they play lacrosse decrease OUTPUT more strongly than others. However, the effect does not reach statistical significance at conventional levels.

Taken together, both alternative proxies show the same picture emerging when proxying non-pecuniary motivation with time elapsed in the pre-treatment stage: The performance reduction is particularly pronounced among motivated workers.

One might also ask why we do not employ pre-treatment work output as a measure for non-pecuniary motivation. We did not pre-register work output for identifying motivation since performance is likely a noisy measure, depending not only on motivation, but also on skills, cognitive ability, experience, luck and other confounding factors. If performance is indeed a noisy measure, we should observe and face a regression-to-the-mean issue. Figure A.5 displays a locally weighted regression of output in the experimental stage against output in the pre-treatment stage. The low performers from stage 1 become better in stage 2 and provide more correct pictures, independent of the treatment group. Also, initial high performers become worse in stage 2 and reduce their output. Thus, we indeed document substantial regression to the mean. Note that Controlled group performs worse than the Baseline, and note that there are only very few workers at both extremes (who provide either just a few, or almost all correct pictures).

Figure A.5: Performance in pre-treatment stage vs. performance in experimental stage



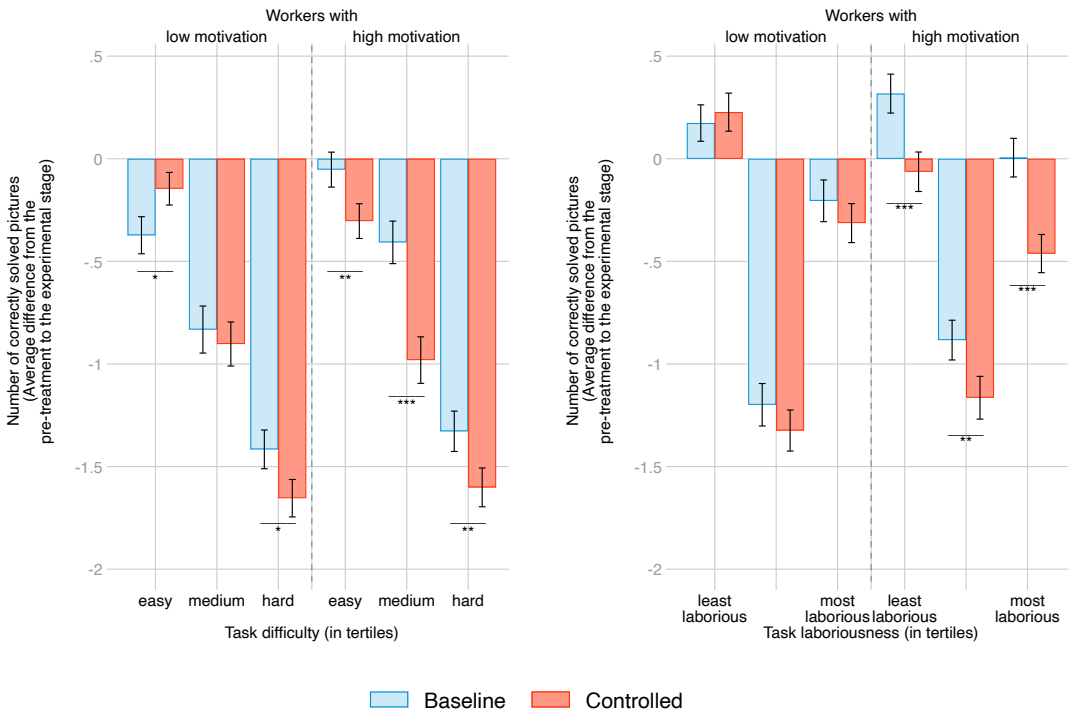
Note: By treatment group, the graph reports a locally weighted regression (default bandwidth) of performance in stage 2 against performance in stage 1. The graph reports on the vertical axis the number of correctly provided pictures in the experimental stage, and on the horizontal axis the number of correctly provided pictures in the pre-treatment stage.

A.2.3 Performance by task heterogeneity and type of worker

Let us revisit Result 2. The performance reduction among complex tasks should be driven by the motivated workforce, too. When splitting the sample by workers' non-pecuniary motivation (see the panel to the right in Figure A.6), we find that Controlled workers with low non-pecuniary motivation actually perform, compared to the Baseline, better in the easy picture category, equally in the medium picture category, and worse among hard-to-solve pictures. In contrast, Controlled workers with high non-pecuniary motivation significantly reduce performance in all pictures categories. The magnitude of the effect amounts to 0.25 pictures or 4.7% among easy pictures ($p < .05$), to 0.57 pictures or 13.3% among medium pictures ($p < .01$) and to 0.27 pictures or 31.5% among challenging pictures ($p < .05$).

The right panel depicts that the performance reduction among the most laborious pictures is due to the motivated workforce. The treatment effect again grows in size with pictures requiring more labor: The performance reduction is with 6.8% smallest among pictures that require the least effort ($p < .01$) and with 16.1% largest among the most time-demanding pictures ($p < .01$).

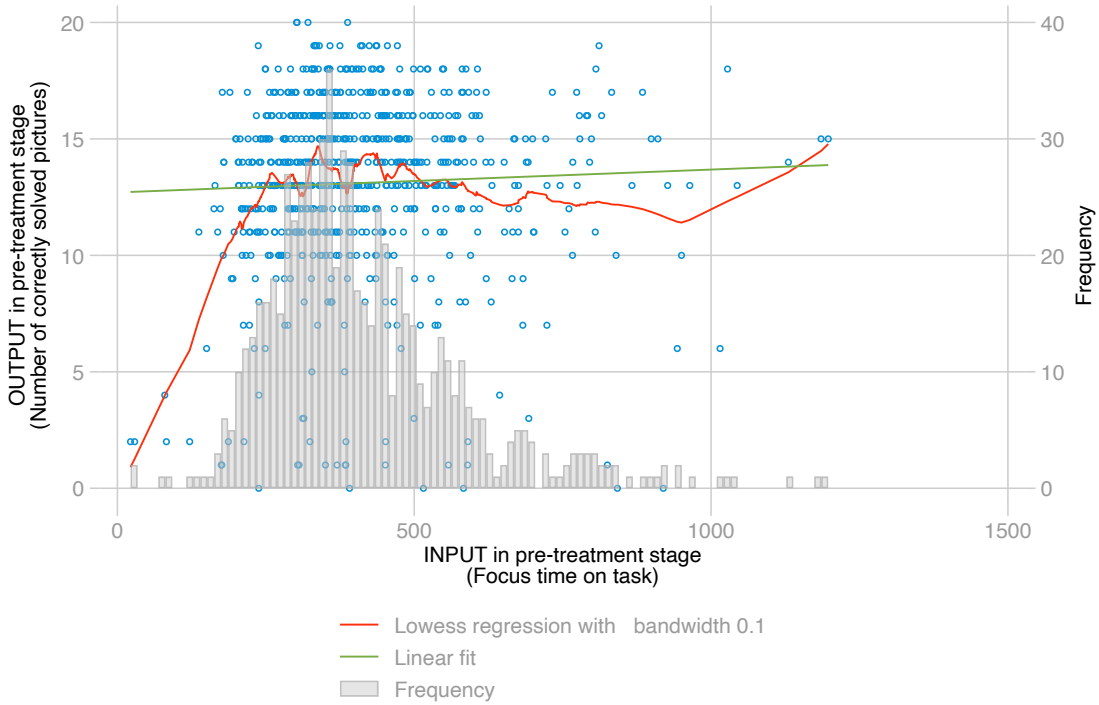
Figure A.6: Performance by task heterogeneity and type of worker



Note: The graph reports on the vertical axis the number of correctly transcribed pictures (OUTPUT) as an average difference from the pre-treatment to the experimental stage, representing the change in performance. The left panel reports the performance difference by task difficulty, the lower panel by task laboriousness. For each stage separately, pictures are classified into difficulty tertiles based on the performance of the Baseline group and into task laboriousness tertiles based on the time elapsed of the Baseline group. Workers are classified into low and high non-pecuniary motivation based on a median split of pre-treatment work input (measured through time on task). Group sizes: $N = 693$. Low non-pecuniary motivation N=346, whereof Baseline n=161, Controlled n=185. High non-pecuniary motivation N=347, whereof Baseline n=189, Controlled=158.

A.2.4 Relationship between performance and time on task (pre-treatment stage)

Figure A.7: Relationship between performance (work OUTPUT) and time on task (work INPUT) in the pre-treatment stage



Note: The graph shows a scatter plot of work OUTPUT (number of correctly solved pictures) on the horizontal axis versus work INPUT (focus time on task) on the vertical axis, all data from stage 1 that is the pre-treatment stage. A histogram of INPUT is overlaid, as well as a the linear fit in green and a lowess regression fit in red.

B The real effort task

B.1 Example Pictures

Figure B.1: Examples of pictures

(a) A blurry picture with incomplete information

(b) An easy-to-solve picture

(c) A picture of medium difficulty

(d) A hard-to-solve picture

B.2 Pre-Treatment Stage

Workers were introduced to the pre-treatment stage in the following way.

A screen shot of the page where workers transcribed the pictures is enclosed in the main body of the paper. Page 4 illustrates an example to help workers understand the instructions. There were two other pages with examples which are omitted due to redundancy.

Figure B.2: The real effort task, stage 1

(a) First page

Task Description

For a one-time project, we need you to extract information out of 20 images. The HIT contains:

- Extract information out of 20 Lacrosse game-play pictures (detailed instructions on next page).
- Once you have completed the HIT, we will grant you automatically a qualification, giving you the possibility to do a second HIT with another set of pictures: you can work on 20 different pictures and get extra money (additional 1 USD).

Reward:

- The HIT reward is set to 1 USD (for the total of the 20 pictures).

This is a one-time job opportunity.

ATTENTION: You must keep the Mturk window open at any time. Do not refresh the browser window, and do not go back to the previous page. You must disable incognito or private mode in your browser in order to work on this HIT.

Next

(b) Second page

Introduction

Instructions

Please carefully read these instructions.

There are 20 images of Lacrosse games. We need you to extract the following information for each of these pictures:

- The jersey number of the player most in the foreground of the picture (that is the player appearing to be closest to the viewer)
- The color of the jersey of the player in the foreground of the picture (light or dark)
- The total number of players in light jersey visible on the picture
- The total number of players in dark jersey visible on the picture
- The number of referees visible on the picture

Note:

- It may be that there is e.g. no referee in the picture. In such cases, please do not leave the respective field empty, but insert a 0.
- DO NOT COUNT players whose head/helmet is cut off the picture (e.g. only legs captured in the photograph).
- DO COUNT players partially or almost fully obscured by other players (unless you can't determine the associated jersey color).
- DO COUNT all players visible on the image, incl. the players on the sidelines.
- In every Lacrosse game, one team must wear light colored jerseys, while the other team wears dark.
- Referees (officials) wear black-and-white jerseys.
- There may be another game (e.g. soccer) going on in the background on another field - ignore that.
- You can open the image in large-scale by clicking on it.

Next

(c) Third page

Introduction

Instructions

There is a “Unclear image, not all info visible”-button. Please click this button if

- the jersey number of the player in the foreground is not visible
- the image is too blurry to identify all information
- for any other reason one or more of the five requested pieces of information cannot be determined

Note: We are well aware that some images are blurry. Also, sometimes the jersey number of the player in the foreground is not visible. Therefore, we prefer that you click the “unclear image”-button over guessing. This is why this button might be the correct response, and, your pay does not depend on which button you click.

The next pages will show you three solved examples.

Next

Figure B.3: The real effort task, stage 1 (cont'd)

(a) Fourth page

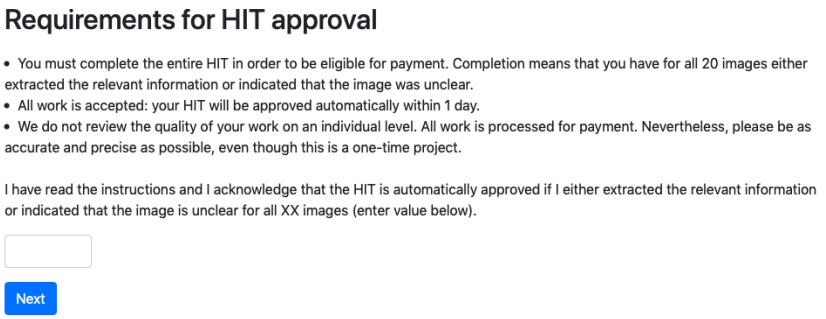
(b) Fifth page

B.3 Experimental Stage

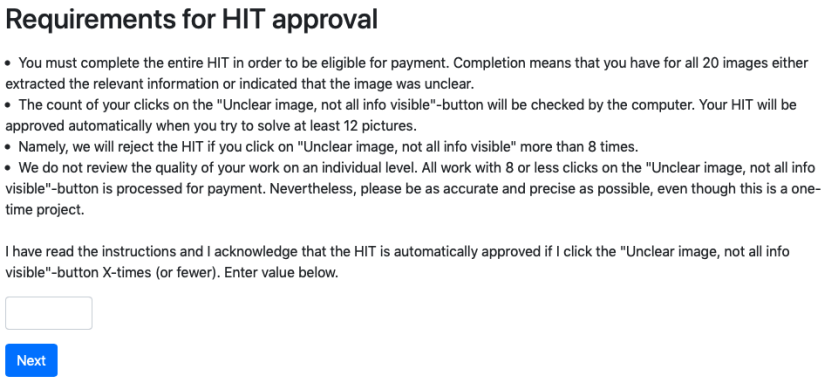
In the experimental stage, workers were already familiar with the task because they completed the pre-treatment stage. Therefore, workers were presented with only two pages: the exact same "Welcome" page as in the pre-treatment stage (refer to figure B.2a) and the page which introduces the treatment, refer to figure B.4a for the Baseline group and to figure B.4b for the Monitored group.

Figure B.4: The real effort task, experimental stage

(a) Instructions for the Baseline group



(b) Instructions for the treatment group Controlled



B.4 Measures

Table B.1: Key Variables

Variable name	Variable type	Dimension	Description	Properties
OUTPUT	outcome	Work output	Number of correctly transcribed pictures, total work output (=20-SKIP-ERROR).	min:0 max:20
SKIP	outcome	Misbehavior	Number of skipped readable pictures.	min:0 max:18
ERRORS	outcome	Misbehavior	Number of transcribed pictures that contain an error.	min:0 max:20

Appendices

C Results Reported Separately by Study

In this section, we report the results of the two trials separately. In general, the qualitative results are very similar. In the first trial (study 1, the original experiment), there is slightly more behavioral heterogeneity in the population compared to the second trial (study 2, the replication). Results that investigate heterogeneous treatment effects are more pronounced in study 1, while average treatment effects are stronger in study 2. In the following, we report all figures and tables that are also reported in the many body of the paper.

C.1 Results of Study 1 (original experiment)

Table C.1: Descriptive statistics, study 1

	Pre-treatment stage		Experimental stage		Difference	
	Baseline	Controlled	Baseline	Controlled	Baseline	Controlled
OUTPUT	13.46	14.15	12.31	12.64	-1.15	-1.51
	(2.97)	(2.81)	(3.43)	(2.53)	(2.73)	(2.45)
SKIP	2.08	1.72	1.83	0.94	-0.25	-0.78
	(1.83)	(1.85)	(2.78)	(1.66)	(2.41)	(1.57)
ERRORS	4.45	4.12	5.86	6.41	1.40	2.29
	(2.34)	(2.40)	(3.35)	(2.31)	(2.92)	(2.46)
Observations	203					

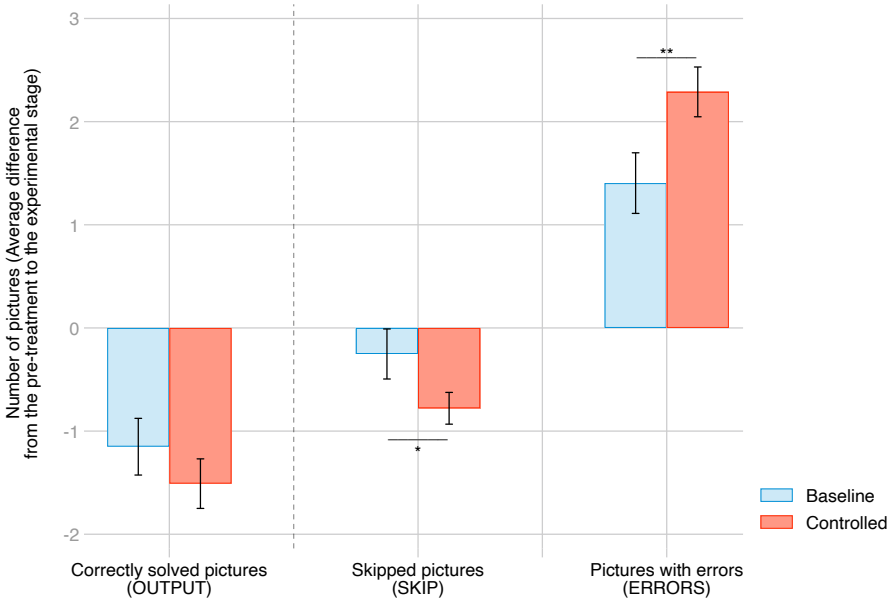
Note: For study 1, the table displays the means along with the associated standard deviation (in parentheses) for the pre-treatment stage, the experimental stage, and the difference between the two stages. Note that workers were randomized into Baseline and Controlled only in the experimental stage. Thus, in the pre-treatment stage, workers were not yet assigned to a group. This implies that workers formed one group in the pre-treatment stage and were only randomly split into Baseline and Controlled in the experimental stage.

C.1.1 Control Reduces Performance

The first result establishes the existence of adverse effects of control.

Result 1. *Control leads to a decrease in average work performance, measured by the count of correctly solved pictures.*

Figure C.1: Average treatment effect on workers' performance, study 1



Note: The graph reports on the vertical axis the number of pictures as an average difference from the pre-treatment to the experimental stage. Errors bars represent the standard error of the mean. The horizontal axis plots work output, representing workers' performance, and its two subdimensions. OUTPUT: Number of correctly solved pictures. SKIP: Number of readable pictures that were declared as unreadable. ERRORS: Number of transcribed pictures that contain an error. $N = 203$, whereof Baseline $n = 99$, Control $n = 104$.
Welch's t-test p values: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Figure C.1 provides support for result 1 and shows that workers in the Baseline on average correctly solve 1.15 fewer pictures in the experimental stage than in the pre-treatment stage. Workers in the Controlled group decrease the number of correctly solved pictures by 1.5. This results in a difference of 0.35 additional unsolved pictures per worker relative to the Baseline. However, this difference is not significant at conventional levels. The reason is that the population in study 1 is quite heterogeneous, as we will later see, and as a consequence, the average treatment effects are neutralized by the two effects that go in the opposite direction.

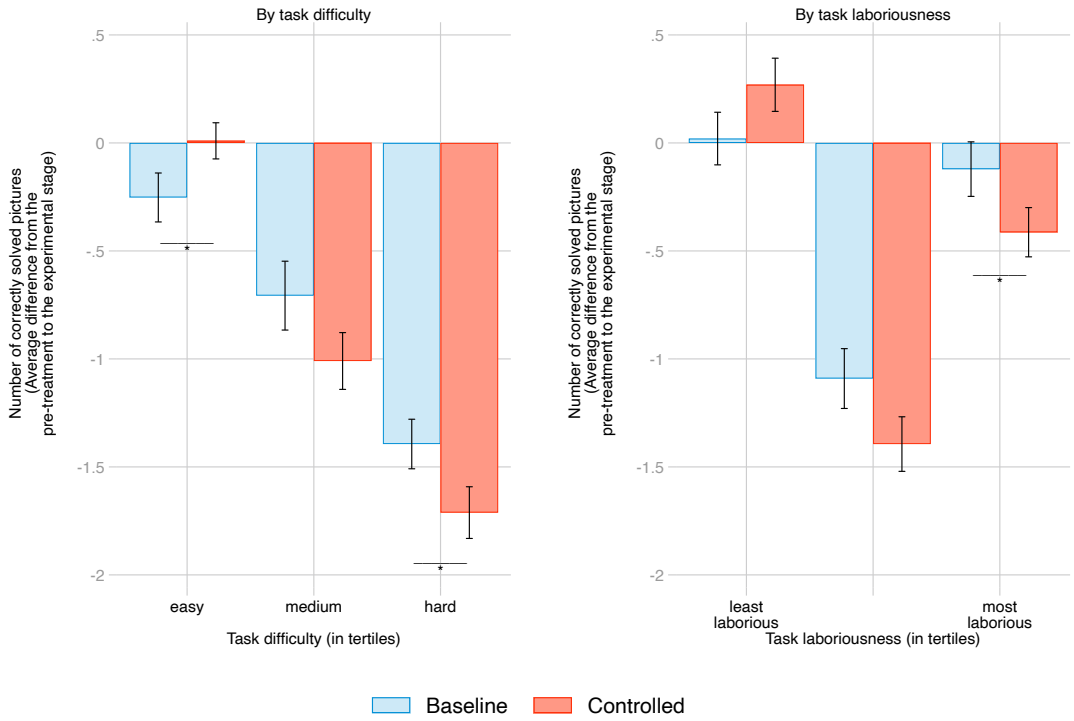
This negative performance effect is due to a significant increase in pictures that contain errors, which is the non-controlled dimension. In the controlled dimension (number of skipped pictures), the control device has a small positive disciplining effect. With regard to the non-controlled dimension, we observe a decline: The number of transcribed pictures that contain errors is significantly lower among controlled workers. Controlled workers submit on average 2.3 more pictures with transcription errors in the experimental stage, while non-controlled workers do so by 1.4 pictures only - a significant difference of 0.9 additional erroneously coded pictures-

Table C.2: Regression Analysis: The effect of the treatment on performance, study 1

	(1) OUTPUT	(2) SKIP	(3) ERRORS
Controlled	-0.11	-0.65	0.75
	(0.35)	(0.27)	(0.36)
OUTPUT (pre-treatment)	0.64		
	(0.09)		
SKIP (pre-treatment)		0.66	
		(0.13)	
ERRORS (pre-treatment)			0.58
			(0.11)
Constant	3.72	0.45	3.26
	(1.24)	(0.27)	(0.53)
r2	0.38	0.31	0.24
N	203	203	203

Note: OLS regressions, robust standard errors (in parentheses). Outcome variables are experimental stage measurements. OUTPUT: Number of correctly solved pictures. SKIP: Number of readable pictures that were declared as unreadable. ERRORS: Number of transcribed pictures that contain an error. Pre-treatment variables of OUTPUT, SKIP and ERRORS control for the level of workers' performance before the treatment was induced.

Figure C.2: Performance by task heterogeneity, study 1



Note: The graph reports on the vertical axis the number of correctly transcribed pictures (OUTPUT) as an average difference from the pre-treatment to the experimental stage, representing the change in performance. The left panel reports the performance difference by task difficulty, the lower panel by task laboriousness. For each stage separately, pictures are classified into difficulty tertiles based on the performance of the Baseline group and into task laboriousness tertiles based on the time elapsed of the Baseline group. $N = 203$, whereof Baseline $n = 99$, Controlled $n = 104$.

C.1.2 Control Reduces Performance Among Challenging Tasks

Result 2. *The negative performance impact of control is significantly more pronounced among hard-to-solve pictures.*

Support for Result 2 is shown in Figure C.2, which plots the average difference of correctly solved pictures by picture difficulty and treatment group. In the left panel, the leftmost bars show that the control device leads to more correct transcriptions of easy-to-solve pictures. Among hard pictures tough, controlled workers perform worse than the Baseline by 0.32 pictures or 24.1% ($p < .10$).

The right panel in Figure C.2 plots a similar graph but by task laboriousness instead of task difficulty: Pictures are ordered into laboriousness tertiles based on the average time spent on a picture in the Baseline group. A similar pattern emerges. We observe that the

performance reduction of controlled workers is especially pronounced among pictures that require more labor. While the performance reduction of the Controlled group compared to the Baseline is not significant among the least and medium laborious pictures, it amounts to and to 0.29 pictures or 12% among the most labor-intensive pictures ($p < .10$).

To asses the robustness of our results, we turn to regression analysis and estimate the models shown in Table C.3.

Table C.3: Regression Analysis: Performance by task heterogeneity, study 1

	(1)	(2)	(3)	(4)	(5)	(6)
	OUTPUT					
	by task difficulty			by task laboriousness		
	easy	medium	hard	least	medium	most
Controlled	0.31	-0.09	-0.28	0.35	-0.16	-0.20
	(0.13)	(0.20)	(0.15)	(0.13)	(0.17)	(0.16)
OUTPUT (pre-treatment)	0.51	0.54	0.64	0.30	0.46	0.65
	(0.16)	(0.09)	(0.06)	(0.10)	(0.07)	(0.06)
Constant	2.46	1.41	-0.46	3.59	1.12	0.67
	(0.94)	(0.46)	(0.15)	(0.57)	(0.34)	(0.16)
r2	0.16	0.18	0.42	0.14	0.18	0.37
N	203	203	203	203	203	203

Note: OLS regressions, robust standard errors (in parentheses). Outcome variables are the experimental stage measurements of the number of correctly solved pictures (OUTPUT) by task difficulty and by task laboriousness, respectively. The 18 readable pictures are classified into three categories by task difficulty based on the number of correctly solved pictures and into three categories by task laboriousness based on the time spent on a picture. The specification controls for the level of workers' pre-treatment performance (OUTPUT) in the respective category.

Column (1) to (3) report the regression coefficients when pictures are classified into three categories based on their difficulty. In the easy picture category (1), controlled workers perform actually better than Baseline workers. The performance reduction occurs among the hard pictures (column (3)). This confirms Result 2: The control device reduces performance in the hard picture category by 0.28 pictures ($p < .10$), conditional on the pre-treatment performance. Again, similar results emerge when we order pictures according to task laboriousness. Controlled workers reduce performance by 0.16 pictures among the medium laborious category and by 0.20 pictures among the labor-intensive tasks.

Taken together, control decreases performance of workers among the most challenging pictures.

Figure C.3: Performance by type of worker, study 1

Note: The graph reports on the vertical axis the number of pictures as an average difference from the pre-treatment to the experimental stage. Errors bars represent the standard error of the mean. The horizontal axis plots work output, representing workers' performance, and its two sub-dimensions. OUTPUT: Number of correctly solved pictures. SKIP: Number of readable pictures that were declared as unreadable. ERRORS: Number of transcribed pictures that contain an error. Workers are classified into low and high non-pecuniary motivation based on a median split of pre-treatment work input (measured through time on task). Group sizes: Low non-pecuniary motivation $N = 101$, whereof Baseline $n = 43$, Controlled $n = 58$. High non-pecuniary motivation $N = 102$, whereof Baseline $n = 56$, Controlled $n = 46$. Welch's t-test p values: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

C.1.3 Control Reduces Performance Among Workers with Non-pecuniary Motivation

As formulated in Hypothesis 3, we expect the performance reduction to be primarily the consequence of a performance reduction by workers with high non-pecuniary motivation when control was absent. Our findings are summarized in result 3.

Result 3. *The negative performance impact of control is significantly more pronounced among workers with high non-pecuniary motivation.*

Support for Result 3 can be seen in Figure C.3 displaying the number of correctly solved pictures and provides evidence supporting result 3: Whereas motivated workers in the Baseline reduce their output by approximately 0.8 correctly solved pictures, motivated workers in the Controlled treatment reduce output by more than 1.9 correctly solved pictures, a significant difference of more than 1 picture ($p < .05$). For workers with low

non-pecuniary motivation, we find no statistically significant differences. In particular, the negative effect of control on motivated workers is significantly stronger than the negative effect of control on workers with low motivation ($p < .05$).

Figure C.3 also displays the number of readable pictures that were declared as unreadable. We do not observe a heterogeneous reaction in the controlled dimension conditional on non-pecuniary motivation. When looking at the non-controlled task dimension, namely the number of pictures that were transcribed erroneously, we find that in the experimental stage, motivated workers in the Controlled treatment increase the number of pictures that contain errors by 2.4. Yet, motivated workers in the Baseline do so only by 1 picture. The difference is highly significant and of substantial magnitude ($p < .01$).

We turn to regression analysis and regress our outcome variables of interest on non-pecuniary motivation as a continuous variable. The results are shown in Table C.4 and confirm the analysis in the previous paragraph: The higher the non-pecuniary motivation of a worker, the stronger the negative reaction to control in our data.

Table C.4: Regression Analysis: Non-pecuniary motivation interacted with treatment, study 1

	(1) OUTPUT	(2) SKIP	(3) ERRORS
Controlled	2.37	-1.40	-0.97
	(0.82)	(0.69)	(0.90)
Non-pecuniary motivation	0.31	-0.09	-0.21
	(0.08)	(0.07)	(0.10)
Controlled × Non-pecuniary motivation	-0.39	0.12	0.27
	(0.12)	(0.09)	(0.13)
OUTPUT (pre-treatment)	0.62		
	(0.09)		
SKIP (pre-treatment)		0.65	
		(0.13)	
ERRORS (pre-treatment)			0.59
			(0.11)
Constant	1.96	1.03	4.60
	(1.22)	(0.60)	(0.84)
R^2	0.41	0.32	0.26
N	203	203	203

Note: OLS regressions, robust standard errors (in parentheses). Outcome variables are experimental stage measurements. OUTPUT: Number of correctly solved pictures. SKIP: Number of readable pictures that were declared as unreadable. ERRORS: Number of transcribed pictures that contain an error. Non-pecuniary motivation is captured by work input in the pre-treatment stage, measured through time on task (in minutes). Pre-treatment variables of OUTPUT,SKIP and ERRORS control for the level of workers' performance before the treatment was induced.

C.2 Results of study 2 (the repetition)

Table C.5: Descriptive statistics, study 2

	Pre-treatment stage		Experimental stage		Difference	
	Baseline	Controlled	Baseline	Controlled	Baseline	Controlled
OUTPUT	12.61	13.03	11.92	11.53	-0.68	-1.50
	(4.38)	(3.94)	(4.13)	(3.85)	(2.72)	(2.48)
SKIP	2.68	2.32	2.07	1.52	-0.61	-0.80
	(3.17)	(2.64)	(3.10)	(2.38)	(2.08)	(1.52)
ERRORS	4.71	4.65	6.00	6.95	1.29	2.30
	(3.67)	(3.33)	(3.44)	(3.53)	(2.95)	(2.63)
Observations	490					

Note: For study 2, the table displays the means along with the associated standard deviation (in parentheses) for the pre-treatment stage, the experimental stage, and the difference between the two stages. Note that workers were randomized into Baseline and Controlled only in the experimental stage. Thus, in the pre-treatment stage, workers were not yet assigned to a group. This implies that workers formed one group in the pre-treatment stage and were only randomly split into Baseline and Controlled in the experimental stage.

C.2.1 Control Reduces Performance

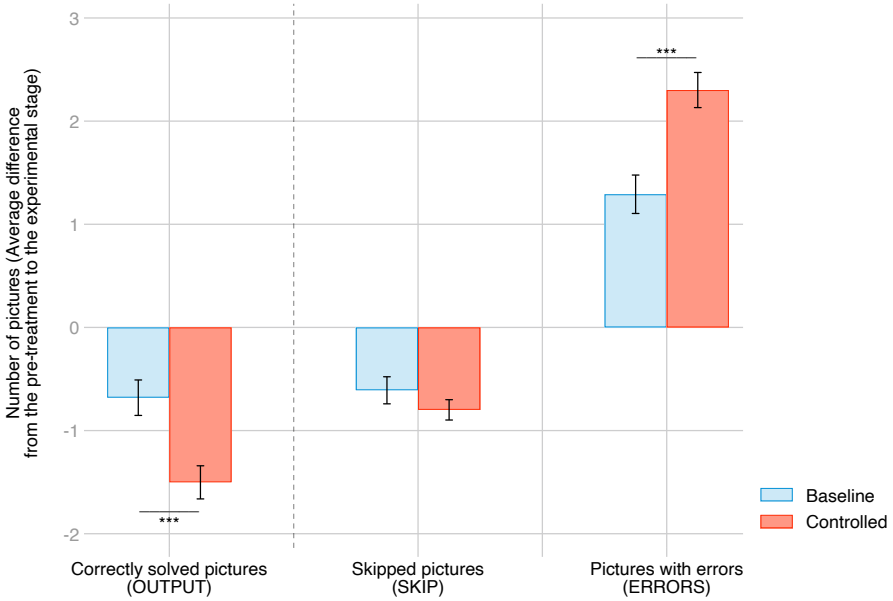
The first result establishes the existence of adverse effects of control.

Result 1. *Control leads to a decrease in average work performance, measured by the count of correctly solved pictures.*

Figure C.4 provides support for Result 1 and shows that workers in the Baseline on average correctly solve 0.7 fewer pictures in the experimental stage than in the pre-treatment stage. Workers in the Controlled group decrease the number of correctly solved pictures by 1.5. This results in a significant difference of 0.8 additional unsolved pictures per worker relative to the Baseline ($p < .01$).

This negative performance effect is due to a significant increase in pictures that contain errors, which is the non-controlled dimension. In the controlled dimension (number of skipped pictures), the controlled device has no significant effect. With regard to the non-controlled dimension, we observe that the number of transcribed pictures that contain errors is significantly higher among controlled workers: Controlled workers submit on average 2.3 more pictures with transcription errors in the experimental stage, while

Figure C.4: Average treatment effect on workers' performance, study 2



Note: The graph reports on the vertical axis the number of pictures as an average difference from the pre-treatment to the experimental stage. Errors bars represent the standard error of the mean. The horizontal axis plots work output, representing workers' performance, and its two subdimensions. OUTPUT: Number of correctly solved pictures. SKIP: Number of readable pictures that were declared as unreadable. ERRORS: Number of transcribed pictures that contain an error. $N = 490$, whereof Baseline $n = 251$, Controlled $n = 239$.
Welch's t-test p values: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

non-controlled workers do so by 1.3 pictures only - a highly significant difference of one additional erroneously coded picture ($p < .01$).

Regression analysis reported in Table C.6 confirms these results.

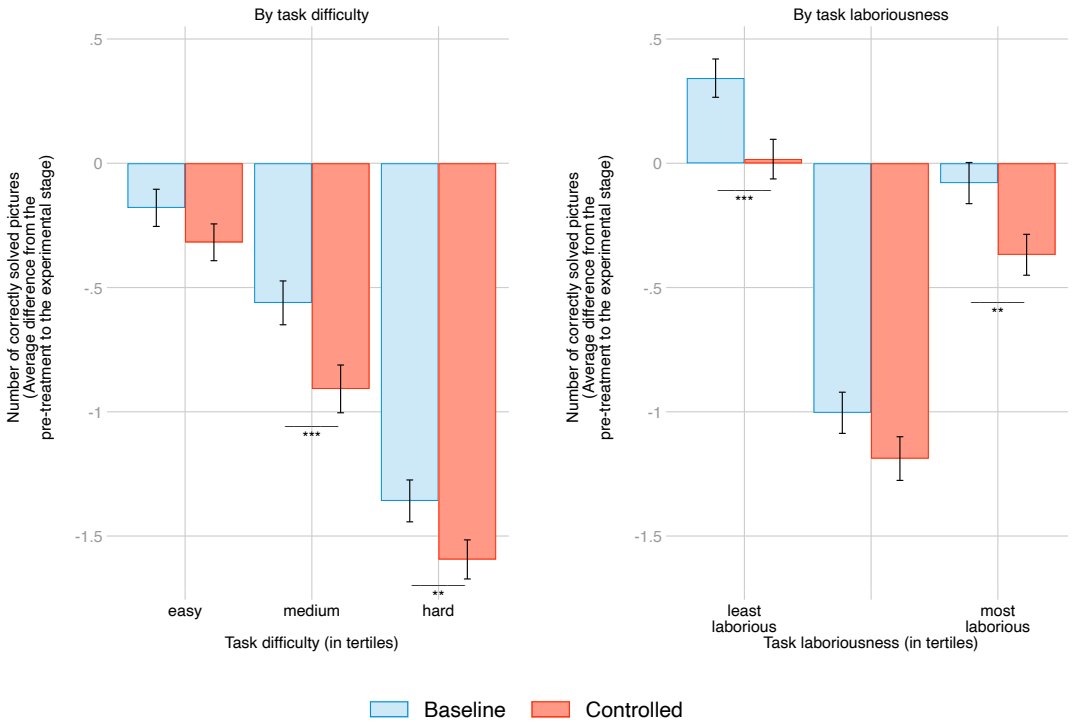
C.2.2 Control Reduces Performance Among Challenging Tasks

We now turn to our second hypothesis, namely that the performance reduction particularly arises in more challenging tasks Our findings are summarized in result 2.

Result 2. *The negative performance impact of control is significantly more pronounced among hard-to-solve pictures*

Support for Result 2 is shown in Figure C.5, which plots the average difference of correctly solved pictures by picture difficulty and treatment group. In the left panel, the leftmost bars show that the control device hardly affects correct transcriptions of easy-to-solve pictures. In the medium category however, Baseline workers solve 0.6 fewer pictures in the experimental stage than in the pre-treatment stage, while Controlled work-

Figure C.5: Performance by task heterogeneity, study 2



Note: The graph reports on the vertical axis the number of correctly transcribed pictures (OUTPUT) as an average difference from the pre-treatment to the experimental stage, representing the change in performance. The left panel reports the performance difference by task difficulty, the lower panel by task laboriousness. For each stage separately, pictures are classified into difficulty tertiles based on the performance of the Baseline group and into task laboriousness tertiles based on the time elapsed of the Baseline group. $N = 490$, whereof Baseline $n = 251$, Controlled $n = 239$.

Table C.6: Regression Analysis: The effect of the treatment on performance, study 2

	(1) OUTPUT	(2) SKIP	(3) ERRORS
Controlled	-0.72	-0.28	0.99
	(0.22)	(0.15)	(0.23)
OUTPUT (pre-treatment)	0.76		
	(0.03)		
SKIP (pre-treatment)		0.75	
		(0.05)	
ERRORS (pre-treatment)			0.68
			(0.05)
Constant	2.31	0.05	2.82
	(0.44)	(0.15)	(0.25)
r2	0.64	0.64	0.47
N	490	490	490

Note: OLS regressions, robust standard errors (in parentheses). Outcome variables are experimental stage measurements. OUTPUT: Number of correctly solved pictures. SKIP: Number of readable pictures that were declared as unreadable. ERRORS: Number of transcribed pictures that contain an error. Pre-treatment variables of OUTPUT, SKIP and ERRORS control for the level of workers' performance before the treatment was induced.

ers solve 0.9 fewer pictures. Controlled workers thus perform worse than the Baseline by 0.3 pictures or 8.8% ($p < .01$). Among hard pictures, this treatment effect grows in magnitude. Controlled workers perform worse compared to the Baseline by 0.24 pictures, which represents a substantial performance reduction of 18.8% ($p < .05$).

The right panel in Figure 5 plots a similar graph but by task laboriousness instead of task difficulty: Pictures are ordered into laboriousness tertiles based on the average time spent on a picture in the Baseline group. A similar pattern emerges. We observe that the performance reduction of controlled workers is especially pronounced among pictures that require more labor. While the performance reduction of the Controlled group compared to the Baseline amounts to 0.33 pictures or 3.6% in the least laborious category ($p < .01$), it amounts to 0.29 pictures or 13% among the most labor-intensive pictures ($p < .05$).

To asses the robustness of our results, we turn to regression analysis and estimate the models shown in Table C.7.

Column (1) to (3) report the regression coefficients when pictures are classified into three categories based on their difficulty. In the easy picture category (1), controlled workers do not perform worse than Baseline workers. The performance reduction occurs among the medium (column (2)) and hard pictures (column (3)). This confirms Result 2:

Table C.7: Regression Analysis: Performance by task heterogeneity, study 2

	(1)	(2)	(3)	(4)	(5)	(6)
	OUTPUT					
	by task difficulty			by task laboriousness		
	easy	medium	hard	least	medium	most
Controlled	-0.13	-0.29	-0.22	-0.22	-0.17	-0.28
	(0.10)	(0.12)	(0.10)	(0.10)	(0.11)	(0.11)
OUTPUT (pre-treatment)	0.94	0.70	0.57	0.68	0.59	0.63
	(0.07)	(0.03)	(0.04)	(0.04)	(0.03)	(0.03)
Constant	0.13	0.74	-0.26	1.80	0.62	0.77
	(0.40)	(0.17)	(0.10)	(0.23)	(0.15)	(0.10)
r2	0.44	0.42	0.35	0.49	0.37	0.40
N	490	490	490	490	490	490

Note: OLS regressions, robust standard errors (in parentheses). Outcome variables are the experimental stage measurements of the number of correctly solved pictures (OUTPUT) by task difficulty and by task laboriousness, respectively. The 18 readable pictures are classified into three categories by task difficulty based on the number of correctly solved pictures and into three categories by task laboriousness based on the time spent on a picture. The specification controls for the level of workers' pre-treatment performance (OUTPUT) in the respective category.

The control device reduces performance in the medium picture category by 0.29 pictures ($p < .05$) and in the hard picture category by 0.22 pictures ($p < .05$), conditional on the pre-treatment performance. Again, similar results emerge when we order pictures according to task laboriousness. Workers do not differ among the medium time-demanding pictures. Controlled workers reduce performance by 0.28 pictures among the most labor-intensive tasks ($p < .01$).

Taken together, control decreases performance of workers among challenging pictures.

C.2.3 Control Reduces Performance Among Workers with Non-pecuniary Motivation

As formulated in hypothesis 3, we expect the performance reduction to be primarily the consequence of a performance reduction by workers that were motivated when control was absent. Our findings are summarized in result 3.

Result 3. *The negative performance impact of control is significantly more pronounced among workers with high non-pecuniary motivation.*

Support for result 3 can be seen in Figure C.6 displaying the number of correctly solved pictures and provides evidence supporting the first part of result 3: Whereas motivated workers in the Baseline reduce their output by approximately 0.35 correctly solved pictures, motivated workers in the Controlled treatment reduce output by more than

Figure C.6: Performance by type of worker, study 2

Note: The graph reports on the vertical axis the number of pictures as an average difference from the pre-treatment to the experimental stage. Errors bars represent the standard error of the mean. The horizontal axis plots work output, representing workers' performance, and its two sub-dimensions. OUTPUT: Number of correctly solved pictures. SKIP: Number of readable pictures that were declared as unreadable. ERRORS: Number of transcribed pictures that contain an error. Workers are classified into low and high non-pecuniary motivation based on a median split of pre-treatment work input (measured through time on task). Group sizes: Low non-pecuniary motivation $N = 245$, whereof Baseline $n = 118$, Controlled $n = 127$. High non-pecuniary motivation $N = 245$, whereof Baseline $n = 133$, Controlled $n = 112$. Welch's t-test p values: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

1.5 correctly solved pictures, a highly significant difference of more than 1 picture. The means are significantly different at the 0.1%-level. For workers with low non-pecuniary motivation, we find no statistically significant differences.

The bars in the middle displays the number of readable pictures that were declared as unreadable. We do not observe a heterogeneous reaction in the controlled dimension conditional on non-pecuniary motivation. The rightmost bars depict the non-controlled task dimension, namely the number of pictures that were transcribed erroneously: In the experimental stage, motivated workers in the Controlled treatment increase the number of pictures that contain errors by 2.3. Yet, motivated workers in the Baseline do so only by 0.9 pictures. The difference is highly significant and of substantial magnitude ($p < .01$). In short, motivated workers significantly reduce the performance, and this is primarily happening in the non-controlled performance dimension.

We turn to regression analysis and regress our outcome variables of interest on non-

pecuniary motivation as a continuous variable. The results are shown in Table C.8. Column (1) reports regressions on the number of correctly solved pictures. It can be seen that the coefficient on the interaction term between the Controlled group dummy and non-pecuniary motivation is negative and statistically significant, again providing evidence that adverse effects of control are primarily occurring among the motivated workforce: The higher the non-pecuniary motivation of a worker, the stronger the negative reaction to control in our data.

Table C.8: Regression Analysis: Non-pecuniary motivation interacted with treatment, study 2

	(1) OUTPUT	(2) SKIP	(3) ERRORS
Controlled	0.27	-0.41	0.13
	(0.57)	(0.45)	(0.60)
Non-pecuniary motivation	0.15	-0.04	-0.10
	(0.06)	(0.04)	(0.06)
Controlled × Non-pecuniary motivation	-0.14	0.02	0.12
	(0.08)	(0.05)	(0.08)
OUTPUT (pre-treatment)	0.76		
	(0.03)		
SKIP (pre-treatment)		0.75	
		(0.05)	
ERRORS (pre-treatment)			0.68
			(0.05)
Constant	1.33	0.35	3.49
	(0.47)	(0.37)	(0.49)
r2	0.64	0.64	0.48
N	490	490	490

Note: OLS regressions, robust standard errors (in parentheses). Outcome variables are experimental stage measurements. OUTPUT: Number of correctly solved pictures. SKIP: Number of readable pictures that were declared as unreadable. ERRORS: Number of transcribed pictures that contain an error. Non-pecuniary motivation is captured by work input in the pre-treatment stage, measured through time on task (in minutes). Pre-treatment variables of OUTPUT,SKIP and ERRORS control for the level of workers' performance before the treatment was induced.