Experimental Economics (2024) 27:469–488
https://doi.org/10.1007/s10683-024-09823-3

Adverse effects of control? Evidence from a field experiment

Holger Herz[1] · **Christian Zihlmann**[2,3]

Received: 1 June 2023 / Revised: 31 January 2024 / Accepted: 3 February 2024 /
Published online: 28 May 2024

Abstract
We conduct a field experiment with Amazon Mechanical Turk ("AMT") workers to causally assess the effect of introducing a control mechanism in an existing work relationship on workers' performance on tasks of varying difficulty. We find that introducing control significantly reduces performance. This reduction occurs primarily on challenging tasks, while performance on simple tasks is unaffected. The negative effects are primarily driven by workers who exhibit non-pecuniary motivation in the absence of control. Our results show that there are adverse effects of control, and they suggest that these adverse effects are of particular concern to firms that rely on high performance on challenging tasks.

Keywords Control · Remote work · Experiment · Crowding out

JEL Classification C93 · D21 · J24 · M5

1 Introduction

While agency theory suggests that monitoring and control are effective in counteracting employee misbehavior (Alchian & Demsetz, 1972; Jensen & Meckling, 1976), some authors have argued that the effect of control may be ambiguous due to a potential reduction in non-pecuniary motivation (Frey, 1993; Frey & Oberholzer-Gee, 1997). Empirical evaluations of control mechanisms in the field are still

✉ Holger Herz
holger.herz@unifr.ch

Christian Zihlmann
christian.zihlmann@bfh.ch

[1] University of Fribourg, Boulevard de Pérolles 90, CH-1700 Fribourg, Switzerland

[2] University of Fribourg, Fribourg, Switzerland

[3] Bern University of Applied Sciences, Bern, Switzerland

limited, and thus gaining a better understanding of the effects of control on worker performance is fundamentally important.

In this paper, we advance our understanding of potential negative effects of control along two dimensions: i) we can assess heterogeneity in reactions to control in the field among workers with different levels of non-pecuniary motivation, and ii) we can causally assess the incidence of potential negative effects across tasks of different difficulty. Understanding such heterogeneity is crucial because task difficulty is often related to the marginal value of a task to the employer. Our setting allows us to study the effects of non-pecuniary motivation and task difficulty independently, whereas they are usually intertwined in observational data.

Our experiment mimics the *introduction* of a control mechanism in an existing work relationship. Specifically, we conduct a pre-registered field experiment on Amazon Mechanical Turk ("AMT"), which is an online crowdsourcing labor market on which employers can recruit workers to perform short jobs for payment.[1] The experiment consists of a pre-treatment and an experimental stage. In the pre-treatment stage, workers receive a flat wage for extracting information from 20 pictures. The work process for each picture consists of two steps: First, workers must declare whether the picture is readable, i.e., whether they can extract the required information. Second, if the picture is declared as readable, workers must extract information from the picture according to the coding guidelines provided. If a picture is declared as unreadable, workers skip the second step of the work process. The pictures vary in difficulty. While some are easy to categorize and require minimal effort, others are more challenging and demanding.

In the experimental stage, workers again face a set of 20 pictures and are randomly assigned to either the "Baseline" or the treatment group ("Restricted"). Conditions in the Baseline are identical to the pre-treatment stage. In Restricted, however, we communicate that we control the number of pictures that are reported as unreadable and implement a maximum allowance threshold: If workers declare more than 8 out of the 20 pictures as unreadable, they will not receive the payment.

In principle, the introduction of a control mechanism in the experimental stage can have two conflicting effects. First, a disciplining effect that increases worker performance by limiting the opportunities for workers in Restricted to shirk by declaring readable pictures as unreadable. Second, the implementation of control could also have detrimental effects on performance if some workers are motivated to perform in the employer's interest even when explicit performance incentives are weak and control is absent (Deci, 1971). Such non-pecuniary motivation could stem from, for example, gift exchange and reciprocity (Akerlof & Yellen, 1990), an individual's desire to perform the task for its own sake (Bénabou & Tirole, 2003), a social norm (Sliwka, 2007), or pride and selfesteem (Ellingsen & Johannesson, 2008). Since the

[1] While AMT is frequently utilized in academia for scientific experiments (Snowberg & Yariv, 2018), its main purpose is a "crowdsourcing marketplace that makes it easier for individuals and businesses to outsource their processes and jobs to a distributed workforce who can perform these tasks virtually" (www.mturk.com, last accessed: 14 November 2023). Workers recruited by us were not aware that the job they are participating in was part of an experiment, and the job is a typical type of task crowd-sourced on AMT.

goal of this paper is to investigate the possible existence of a negative behavioral effect, we deliberately implemented a control mechanism that is as ineffective as possible in disciplining workers: workers who want to shirk can simply declare pictures as readable, but enter false information in the second work step.

Our first result shows that average performance is reduced when control is introduced. The average worker in Restricted reduces performance significantly by 5.5 percent relative to the counterfactual. In addition, the introduction of control reduces both the number of low and high performers, meaning that the variance of worker performance is significantly lower in Restricted than in the Baseline.

Second, we find that the reduction in performance in the Restricted treatment is particularly pronounced among workers who were motivated to perform in the pre-treatment stage, where control was absent. Pre-treatment motivation is measured by the time spent: those workers who invested relatively more time into solving the job are classified as being more motivated. Splitting our sample at median pre-treatment motivation, we find that output among workers with high motivation is reduced by 8.7% in Restricted relative to the counterfactual. In contrast, Restricted workers with low motivation do not reduce performance compared to Baseline workers. Two alternative proxies for non-pecuniary motivation—(i) whether workers have a potential intrinsic interest in the content of the coded pictures and (ii) whether workers re-consulted the coding guidelines while working in the pre-treatment stage—confirm these results. Thus, the implementation of control reduces performance especially among those workers who were motivated to perform in the absence of control, which confirms previous laboratory findings.

Finally, Restricted workers reduce performance particularly among difficult and time-demanding tasks. Compared to the counterfactual, Restricted workers reduce performance by 20.5% among the hardest tertile of pictures. We find a smaller performance reduction of 8.3% in the medium category and no significant difference for the easiest tertile. Similar results are found when sorting the pictures according to laboriousness, defined as the average time spent on a picture. Restricted workers perform significantly worse among the more time-demanding pictures, reducing correct transcription rates by 12.7%. Again, the decrease in worker performance is smaller among the medium (7.2%) and the least labor-intensive pictures (3.0%).

The finding that control differentially affects performance conditional on task difficulty may have important implications, as it suggests that the adverse effects of control are contingent on the value a firm attaches to difficult tasks. In some organizations, the value derived from solving a task may be uncorrelated or even negatively correlated with its difficulty. As a result, performance is reduced on those tasks where it has the least negative impact on the employer.

In other work environments, however, task difficulty is likely to be positively correlated with the value created for the employer. For example, when tasks are complementary inputs in production, those that are less frequently correctly provided (i.e., the difficult ones) tend to have a higher marginal value to the firm (Kremer, 1993). The heterogeneity in performance reductions by task difficulty implies that the average treatment effect may strongly underestimate the negative impact of control on the firm's value of production.

Our findings shed light on the heterogeneity in the use of monitoring and control mechanisms across different work environments (Ichniowski et al., 1997). In many jobs, workers have private information about the importance of different tasks for firm productivity, and firms cannot install monitoring technology that accounts for this private information (Ichniowski & Shaw, 2003; Bartling et al., 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.

There may be good reasons not to implement control in such settings, as our evidence shows. Control has a negative impact on performance, particularly for difficult and tedious tasks. If tasks are complementary inputs into production, or if task difficulty and the marginal value of a task are positively correlated, our data suggests that control could be highly detrimental for the firm. This result aligns well with theoretical predictions that posit the adverse effects of managerial interventions to be particularly pronounced when there are complementarities in production (Friebel & Schnedler, 2011).

This paper contributes to the broader literature on adverse effects of control, monitoring and surveillance, where it has been shown that formal control can impact an agent's trust in the principal and may reduce effort (Falk & Kosfeld, 2006). Laboratory studies have assessed how control can have heterogeneous effects across agent types (Dickinson & Villeval, 2008; Schnedler & Vadovic, 2011; Ziegelmeyer et al., 2012; Maas & Van Rinsum, 2013; Masella et al., 2014; Kessler & Leider, 2016; Riener & Wiederhold, 2016; Burdin et al., 2018; Schmelz & Ziegelmeyer, 2020). Our field experimental evidence gives credence to these findings from the lab, and extends the analysis by showing heterogeneous effects across task difficulty.

The few (quasi-)experimental studies on the effects of control that have been conducted in the field have also been limited to tasks that do not differ in difficulty. When introducing monitoring in unidimensional tasks, Nagin et al. (2002) find that lowering the level of monitoring leads most workers to decrease performance. Similarly, implementing monitoring has been found to decrease employee theft (Pierce et al., 2015), and to increase worker performance when tasks are unidimen sional (Boly, 2011). Belot and Schröder (2016) investigate the effects of monitoring in a multidimensional job and find that monitoring increases performance in the monitored dimension, but decreases performance in the non-monitored dimension. We go beyond these articles by studying heterogeneous treatment effects conditional on task difficulty. Table A.1 in the Online Appendix provides a concise overview of the literature.

2 The experiment

2.1 The real effort task

The field experiment was conducted on AMT. We recruited workers as a neutral AMT employer and workers were not aware that they participate in a study.

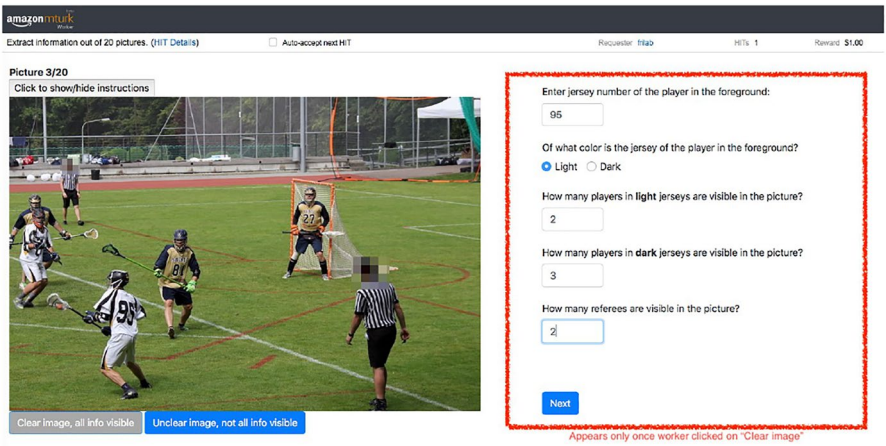

Fig. 1 The real effort task

Workers engage in a visual search task: extracting and categorizing information from a picture.[2] 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.

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).[3] If the picture is declared readable, workers have to enter five pieces of information in a second step. The entry form is shown in Fig. 1.

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 Online Appendix provides examples of pictures of different difficulty). Each worker transcribed the same set of pictures, and the sequence of pictures was randomly determined for each worker by the computer.

[2] The task was programmed with the software oTree (Chen et al., 2016).

[3] 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.

2.2 Set-up and treatments

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 not present and other external incentives are minimized. Workers are truthfully informed that the task is automatically approved and paid regardless of the provided work ("All work is accepted: your job will be approved automatically within 1 day", which is an often used function on AMT, see Online Appendix B for the full instructions). Thus, workers can report all 20 pictures as unreadable and still receive the payment.

This stage has a two-fold purpose. 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 (Horton et al., 2011). Second, it allows us to observe behavior of all participants in an environment without control.

2.2.2 The experimental stage

Once the workers have completed the pre-treatment stage, they are automatically offered the opportunity to do another, different set of 20 pictures. The order of appearance is again randomized by the computer for each worker individually.

If workers accept the offer, they are randomly assigned to one of two groups: In "Baseline", they receive the same contract as before, a flat payment of USD 1, and the job is auto-approved and paid regardless of the work performed.

In "Restricted", however, they are assigned to a control mechanism: Workers are truth-fully informed that they are allowed to declare a maximum of 8 out of 20 pictures as unclear and that this will be controlled and verified automatically by the computer. If workers do not exceed the maximum allowance threshold, a flat reward of USD 1 is automatically paid. If the requirement is not met, workers are not eligible to receive the payment.

2.3 Measures, procedures and hypotheses

2.3.1 Measures

To produce a correct transcription of a picture, workers first need to identify readable picures as readable. Then, they 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: (i) declaring a picture as unreadable even though it is readable, 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 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 wrongly

transcribed. The variable OUTPUT captures overall work output, measured by the total number of correctly solved pictures (note that there are 20 pictures in total and therefore: OUTPUT = 20 – SKIP – ERRORS). OUTPUT thus represents worker performance and is our main variable of interest. 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.

2.3.2 Hypotheses

Our first hypothesis concerns the potential negative effect of implementing control on performance in our setting. The control technology used in the Restricted treatment restricts workers' shirking possibilities by limiting the option to declare pictures as unreadable. While workers were technically able to mark more pictures as unreadable than the maximum allowed, the enforcement of this limit was through payment: workers who did not meet the minimum performance requirement were not eligible for payment. However and importantly, the control technology leaves the option open to erroneously and effortlessly transcribe the pictures. Therefore, opportunistic agents can easily bypass the control technology and consequently, we do not expect a disciplining effect. On the other hand, if control is detrimental because workers react negatively to the implementation of control, Restricted workers should reduce performance (Barkma, 1995; Frey, 1993; Frey & Jegen, 2001). Hypothesis 1 thus assesses the external validity of the laboratory finding that controlling workers may backfire (see, for example, Falk & Kosfeld, 2006; Ziegelmeyer et al., 2012; Schmelz & Ziegelmeyer, 2020).

Hypothesis 1 Introducing control reduces performance.

Our second hypothesis is concerned with heterogeneity across workers in their behavioral reaction to control. 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 (Akerlof & Yellen, 1990; Bénabou & Tirole, 2003; Ellingsen & Johannesson, 2008; Fehr et al., 1993; Sliwka, 2007). 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 (Dickinson & Villeval, 2008). We thus expect that the decrease in performance in response to the control mechanism will be particularly pronounced among workers with non-pecuniary motivation.

Hypothesis 2 The adverse effect of control is particularly pronounced among workers 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 the job in the pre-treatment stage, as a proxy for non-pecuniary motivation. 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.[4]

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. 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 may be more motivated to engage with our task. 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 job. Workers who re-consult the guidelines are classified as workers with higher non-pecuniary motivation, because they strive to complete the task correctly according to the guidelines provided.

Our third hypothesis assesses heterogeneous reactions to control across types of tasks. Workers are tasked with transcribing pictures that vary 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 Introducing control reduces performance among challenging tasks. The adverse effect of control is particularly pronounced among the hard-to-solve pictures.

[4] 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. 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 $\rho = 0.09$, $p = 0.02$), 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 perform better in stage 2. The opposite holds true for high performers. See Figure C.6 in the Online Appendix.

2.3.3 Procedures

We conducted two randomized control trials, the first on December 10th, 2018 and the second from March 9th to 11th, 2020.[5] Both trials were pre-registered before data collection. We conducted a second trial because only a subset of our empirical analyses were pre-registered before the first trial. In the following, we highlight those hypotheses for which adjustments in the pre-analysis plan were made between trial 1 and trial 2. Hypothesis 1 was pre-registered in both the analysis plans of study 1 and study 2. Hypothesis 2 was pre-registered in both the analysis plans of study 1 and study 2, too. However, 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. We report results for both proxies. Hypothesis 3 was pre-registered for the second trial, after exploratory findings in the first trial.

The total sample consists of 693 workers, 203 workers in the first trial and 490 in the second.[6] 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 only once they started the experimental stage. 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 Restricted group. We thus deem attrition to be low and not significantly differently distributed across treatments. Moreover, dropped out workers do not exhibit significant differences among any of the three performance dimensions depending on treatment assignment.

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 min for each stage, yielding an hourly pay of approximately USD 9.

[5] 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 Online Appendix D.

[6] In the first trial, 221 workers completed the experimental stage. 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.

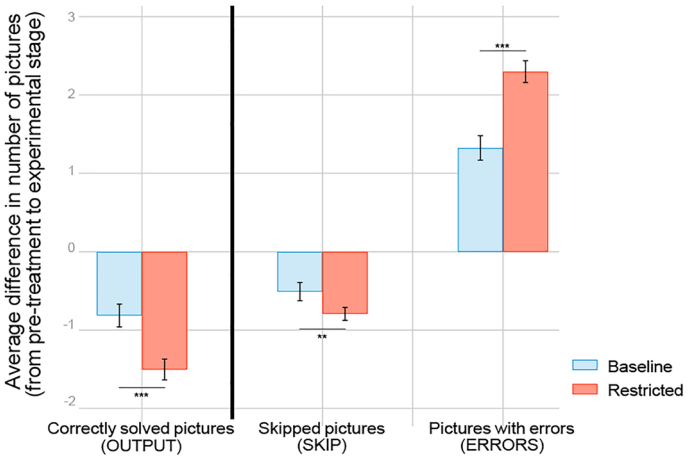

Fig. 2 Average treatment effect on workers' performance. *Note*: The graph reports on the vertical axis the number of pictures as an average difference from the pre-treatment to the experimental stage. Errors bars represent the standard error of the mean (accounting for unequal variances). OUTPUT: Number 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 $n = 350$, Restricted $n = 343$. Unequal variance t-test p values: $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$

3 Results

In the main body of the paper, we report our results based on the difference in the outcome variable between the experimental and the pre-treatment stage. In the Online Appendix, we additionally provide analyses based on a regression approach by investigating the experimental stage outcomes conditional on the pre-treatment measurements.[7] Descriptive statistics of all main outcome measures are presented in Table C.1 in the Online Appendix.

3.1 Control decreases worker performance

Our first result establishes the existence of adverse effects of control in our setting.

Result 1 Introducing control leads to a significant decrease in average work performance.

Figure 2 provides support for Result 1.[8] It shows that workers in the Baseline on average solve 0.8 fewer pictures correctly in the experimental stage than in the pre-treatment stage (variable OUTPUT). Notably, Restricted workers decrease

[7] 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).

[8] All reported results are computed with Stata (StataCorp, 2019), using the graphical schemes of Bischof (2017).

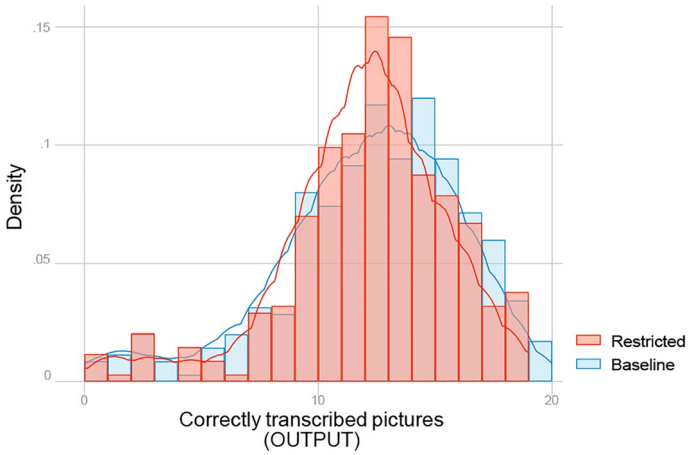

Fig. 3 Histogram and kernel density estimates of workers' performance. *Note*: The graph reports by experimental group a histogram of the variable OUTPUT (number of correctly transcribed pictures). The data are experimental stage measurements. The bin width is set to 1 because the data is discrete. Epanechnikov kernel density estimates are overlaid, the default (optimal) width was used

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 ($p < 0.01$).[9] This is equivalent to 5.5% reduction in output compared to the counterfactual situation in which workers in Restricted had remained uncontrolled.[10]

We test the robustness of this result by regressing experimental stage measurements on the treatment dummy while conditioning on the pre-treatment stage measurements to control for individual pre-treatment characteristics. We again find that the introduction of control reduces performance by 0.56 correctly solved pictures ($p < 0.01$, see Table C.2 in the Online Appendix).

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 Restricted treatment in the experimental stage (Figures C.1 and C.2 in the Online Appendix display distribution plots of SKIP and ERRORS). The kernel density estimates for Restricted 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

[9] In this subsection, if not otherwise explicitly mentioned, when comparing two groups, we report p values from Satterthwaite'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.

[10] Throughout the paper whenever we report percentage differences, they are calculated using thefollowing formula: $\frac{X_2^T}{X_1^T + \Delta X^B}$, where X_i^T is the stage i variable of interest in the Restricted group, and $\Delta X^B = X_2^B - X_1^B$ is the difference between stage 1 and stage 2 in the Baseline group, such that $X_1^T + \Delta X_2^B$ constitutes the counterfactual change for the Restricted group had they not been controlled.

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 the control mechanism ($p = 0.02$). Put differently, control cultivates the average worker.

Figure 2 also provides insights about the effects of control in the two steps of the work process. Restricted workers reduce the number of skipped readable pictures by 0.8 between the pre-treatment stage and the experimental stage while non-restricted workers do so by 0.5 pictures only ($p = 0.05$). Simultaneously, we observe the number of transcribed pictures with errors to be 16.8% higher among Restricted workers compared to the counterfactual, a highly significant difference ($p < 0.01$). Regression analysis (see columns (2) and (3) in Online Appendix Table C.2) again confirms these findings.

The finding that Restricted workers increase ERRORS relative to the Baseline is robust to applying various alternative measurements for work quality, such as (i) error rates instead of absolute numbers, (ii) errors by single input field instead of full pictures, or (iii) errors by single input field per picture (see Online Appendix C.3). Restricted 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. It is noteworthy that the number of SKIP decreases in the presence of control.

The adverse effects of control arise in the non-restricted work step. 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. Our observation that performance increases in the monitored dimension despite the absence of a binding incentive mechanism is also consistent with recent evidence documenting that simply making monitoring more visible leads to an increase in the monitored performance dimension, even though incentives remain unchanged since workers are not paid for the monitored performance dimension (Jensen et al., 2020). Our finding also aligns well with Anteby and Chan (2018) who show that control may encourage workers to engage in deviant behavior in dimensions that are difficult to detect for the employer.

This implies that it can be difficult for firms to notice the detrimental impact of control, because performance metrics in the controlled dimension are likely to signal positive effects. However, the adverse effects of control may arise in other, potentially non-observed steps of the work process. Thus, performance metrics relating to control mechanisms may be misinterpreted and lead to false conclusions about the effectiveness of control.

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 control or whether there is important heterogeneity in workers' behavioral response.

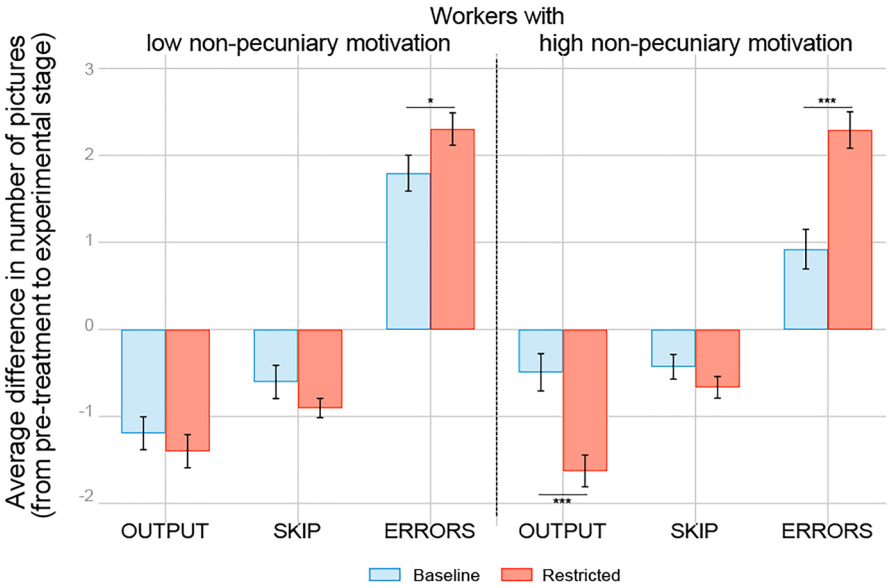

Fig. 4 Performance by type of worker. *Note*: The graph reports on the vertical axis the number of pictures as an average difference from the pre-treatment to the experimental stage. Errors bars represent the standard error of the mean (accounting for unequal variances). OUTPUT: Number of correctly solved 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 $n = 161$, Restricted $n = 185$. High non-pecuniary motivation $N = 347$, whereof Baseline $n = 189$, Restricted $n = 158$

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

Support for Result 2 can be seen in Fig. 4. As explained in Sect. 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.

The right panel provides evidence supporting Result 2: Whereas motivated workers in the Baseline reduce their performance by about 0.5 pictures, motivated workers subject to the control mechanism reduce output by 1.6 pictures, a highly significant difference of more than one picture. This is equivalent to a decrease of output by about 8.7% ($p < 0.01$) when motivated workers are Restricted. For workers with low motivation, depicted in the left panel, we do not find significant differences in output between the two groups. Thus, the negative performance effect of control on motivated workers is significantly stronger than the negative performance effect of control on workers with low motivation ($p = 0.02$). We also observe that

the reduction in work output for motivated workers in the Restricted condition is not because they skip more pictures, but because they make more errors than workers who are not restricted.

To test the robustness of our results, we regress our outcome variables of interest on individual non-pecuniary motivation, measured as time spent on the task, both continuously and via a median split. We find that both interaction terms are negative and highly statistically significant, indicating that workers with high non-pecuniary motivation are those that react especially adverse to the implementation of control (regression results are shown in Table C.3 in the Online Appendix).

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 click the "Open Instructions"-button in the pre-treatment stage to reconsult the instructions on how to classify pictures properly and (ii) whether workers play or regularly watch lacrosse.[11] In case (i), 144 workers re-consulted the guidelines at least once, and are thus classified as motivated. Motivated workers reduce performance by 9.6% when Restricted ($p < 0.01$), while non-motivated workers do so by 4.2% only ($p = 0.02$). A regression reveals that the difference-in-difference interaction term is also marginally significant ($p = 0.07$), see Table A.1 in the Online Appendix. In case (ii), 151 workers play or regularly watch lacrosse and are thus classified as motivated. Motivated workers reduce performance by 8.9% when Restricted ($p = 0.06$), while non-motivated workers do so by 4.7% only ($p < 0.01$). While the difference-in-difference interaction term goes in the right direction, it fails to reach conventional levels of statistical significance.[12]

3.3 Control reduces worker performance among challenging tasks

To assess our third hypothesis, we categorize the 18 readable pictures into three categories based on their difficulty, measured by the achieved performance (OUTPUT), as pre-registered for the second experiment.[13] 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 the introduction of control is significantly more pronounced among hard-to-solve pictures.

Support for Result 3 is shown in Fig. 5, which plots the average difference of correctly solved pictures by picture difficulty and experimental group. In the left

[11] Detailed results of these additional analyses are presented in Online Appendix C.3.1.

[12] 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.

[13] 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.

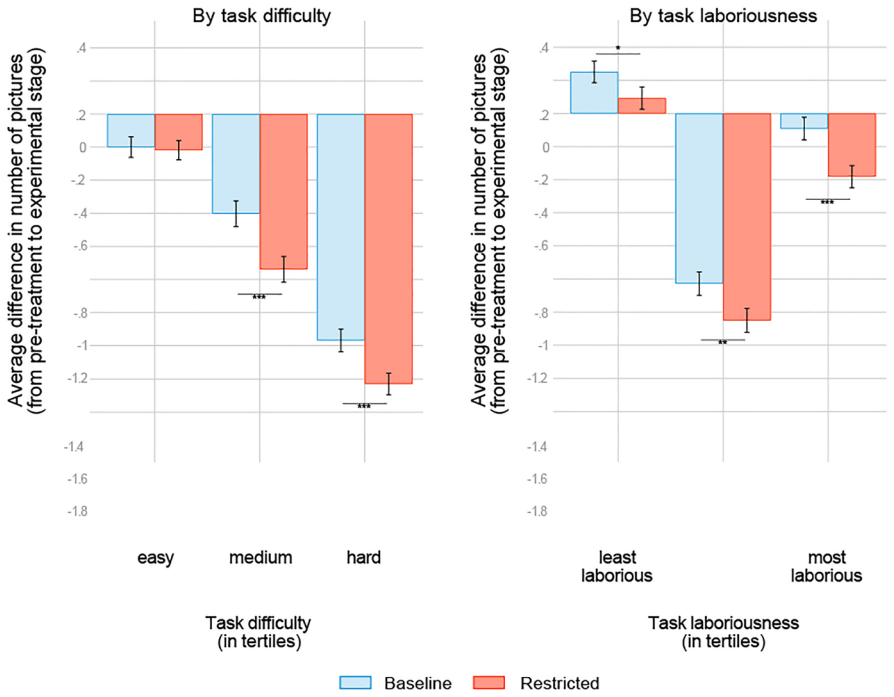

Fig. 5 Performance by task heterogeneity. *Note*: The graph reports on the vertical axis the number of correctly transcribed pictures (OUTPUT) as an average difference from the pre-treatment to the experimental stage, representing the change in performance. The left panel reports the performance difference 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 $n = 350$, Restricted $n = 343$

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 Restricted workers solve 0.9 fewer pictures. Restricted workers thus perform worse than the Baseline by 0.3 pictures, which is an 8.3% reduction in performance ($p < 0.01$). Among hard pictures, this treatment effect grows in relative magnitude. Restricted workers perform worse compared to the Baseline by 0.26 pictures, which represents a substantial 20.5% reduction in performance ($p < 0.01$).

The right panel in Fig. 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. We observe that the relative performance reduction of Restricted workers is especially pronounced among pictures that require more labor. While the performance reduction of Restricted workers compared to non-restricted workers amounts to 0.16 pictures or 3.0% in the least laborious category ($p = 0.09$), it amounts to 0.22

pictures or 7.2% in the medium category ($p = 0.03$) and to 0.29 pictures or 12.7% among the most labor-intensive pictures ($p < 0.01$).

Finally, we again assess the robustness of our results with regression analysis, which confirms the results reported above (see Table C.6 in the Online Appendix). In addition, regression analyses reveal that the performance reduction among hard and labor-intensive tasks is primarily driven by the motivated workforce (see Figure C.7 in the Online Appendix).[14]

Therefore, the average reduction of performance documented in Result 1 can be primarily attributed to workers with non-pecuniary motivation, who reduce their performance particularly among laborious and hard-to-solve tasks.

4 Conclusion

This article provides novel evidence on the adverse effects of control in the field. We document that the introduction of control in an existing work relationship adversely affects worker performance, particularly for difficult and labor-intensive tasks and for workers with non-pecuniary motivations. These results have important implications for the optimal design of control in firms. In particular, they imply that the implementation of control can be profoundly harmful (1) for firms whose workforce is motivated to perform even when extrinsic incentives are largely absent, and (2) for firms that derive particularly high marginal value from worker performance on challenging tasks. For example, in work environments where different tasks are complements, difficult tasks are likely to have the highest marginal value to the firm. The average treatment effect on performance may substantially underestimate the impact of control on firm profitability in such settings.

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 to identify potential negative effects of control. Our results show that moderately effective control mechanisms will likely have positive overall performance effects, in particular when tasks are perfect substitutes.

Together with theoretical work (Bénabou & Tirole, 2003; Ellingsen & Johannesson, 2008; Frey, 1993; Sliwka, 2007), this article conveys an important lesson for firms when implementing monitoring and control technologies. The introduction of these technologies may distort performance in unintended ways, which can be particularly severe when only some relevant dimensions of the work task can be targeted. More research on how organizations can install control technology while avoiding unwanted side effects is warranted.

In this regard, it is important to note that our findings relate to a situation in which control is newly and uniformly implemented within an existing work relationship. Such an implementation of control can be interpreted as a signal of distrust, which may be one potential mechanism that causes the adverse effects of control.

[14] As noted in Sect. 2.3.1, Hypothesis 3 was only pre-registered before trial 2, and here we present the results for our pooled sample. However, in Online Appendix D.2, we perform our analyses using data from trial 2 only, and qualitatively replicate the results presented here.

Indeed, call center workers perceive the implementation of a smartphone ban as a signal of distrust, and they generally prefer trust to control (Chadi et al., 2022). Our results, however, do not necessarily generalize to situations in which workers enter a firm that is already using control technology.

Imposing controls, mandates, and regulation is also highly important in policy making, and there is a growing literature documenting crowding out effects in public policy (Chater & Loewenstein, 2023). For example, mandatory enforcement of policies in the context of Covid-19, such as vaccination mandates, may crowd out citizens' motivation and their voluntary support for these measures (Schmelz, 2021; Schmelz & Bowles, 2021).

Finally, the behavioral heterogeneity in our data has important implications for the design of organizations. Ultimately, how can an organization design incentive schemes that discipline the opportunistic workers without reducing performance of those with non-pecuniary motivations? Moreover, the existence of different control regimes across and within firms raises interesting questions (Beckmann & Kräkel, 2022). For example, the literature documents worker self-selection with respect to other behavioral factors, such as overconfidence (Larkin & Leider, 2012), cooperation (Kosfeld & Von Siemens, 2011) or a preference for being one's own boss (Bartling et al., 2014; Hamilton, 2000; Hurst & Pugsley, 2011). Initial evidence suggests that some workers value flexible work arrangements (Angelici & Profeta, 2023) and are willing to forgo monetary compensation in exchange for not being monitored (Liang et al., 2022). If workers are heterogeneous in their degree of control aversion, we may also see sorting into firms and industries. The (non-) use of control technology may then become a strategic tool for firms to attract specific types of workers.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10683-024-09823-3.

Acknowledgements This experiment was registered before data collection on the AEA RCT registry (ID: AEARCTR-0003475) and approved by the ethics committee of the Internal Review Board of the University of Fribourg, Switzerland (Ref-No.: 393). The authors declare that they have no other relevant or material financial interests that relate to the research described in this paper. We are grateful for valuable comments to Björn Bartling, Berno Büchel, Adrian Chadi, Alain Cohn, Martin Huber, Michael Kosfeld, Ian Larkin, Victor Maas, Marina Schröder, Christian Zehnder, to conference participants at the AEA Annual Meeting in New Orleans 2023, the EEA Annual Congress 2021, the GEABA Symposium 2022, the NCBEE 2019, and the SSES Annual Congress 2021, as well as to seminar participants at the Florida State University, Kandersteg, and the University of Fribourg. The replication material for the study is available at https://doi.org/10.17605/OSF.IO/R9ATB.

Funding Open access funding provided by University of Fribourg.

References

Akerlof, G. A., & Yellen, J. L. (1990). 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.

Angelici, M., & Profeta, P. (2023). Smart working: Work flexibility without constraints. *Management Science*. https://doi.org/10.1287/mnsc.2023.4767

Anteby, M., & Chan, C. K. (2018). A self-fulfilling cycle of coercive surveillance: Workers' invisibility practices and managerial justification. *Organization Science, 29*(2), 247–263.

Barkma, H. (1995). Do top managers work harder when they are monitored? *Kyklos, 48*(1), 19–42.

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). Screening, competition, and job design: Economic origins of good jobs. *American Economic Review, 102*(2), 834–864.

Beckmann, M., & Kräkel, M. (2022). Empowerment, task commitment, and performance pay. *Journal of Labor Economics, 40*(4), 889–938.

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

Bénabou, R., & Tirole, J. (2003). Intrinsic and extrinsic motivation. *The Review of Economic Studies, 70*(3), 489–520.

Bischof, D. (2017). New graphic schemes for stata: Plotplain and plottig. *The Stata Journal, 17*(3), 748–759.

Boly, A. (2011). 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.

Burdin, G., Halliday, S., & Landini, F. (2018). The hidden benefits of abstaining from control. *Journal of Economic Behavior & Organization, 147*, 1–12.

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

Chadi, A., Mechtel, M., & Mertins, V. (2022). Smartphone bans and workplace performance. *Experimental Economics, 25*(1), 287–317.

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.

Chater, N., & Loewenstein, G. (2023). The i-frame and the s-frame: How focusing on individual-level solutions has led behavioral public policy astray. *Behavioral and Brain Sciences, 46*, e147.

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*, 88–97.

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

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.

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

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). 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.

Frey, B. S., & Jegen, R. (2001). Motivation crowding theory. *Journal of Economic Surveys, 15*(5), 589–611.

Frey, B. S., & Oberholzer-Gee, F. (1997). The cost of price incentives: An empirical analysis of motivation crowding-out. *The American Economic Review, 87*(4), 746–755.
Friebel, G., & Schnedler, W. (2011). Team governance: Empowerment or hierarchical control. *Journal of Economic Behavior & Organization, 78*(1–2), 1–13.
Hamilton, B. H. (2000). Does entrepreneurship pay? An empirical analysis of the returns to self-employment. *Journal of Political Economy, 108*(3), 604–631.
Horton, J. J., Rand, D. G., & Zeckhauser, R. J. (2011). 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, 2011*, 73–143.
Ichniowski, C., & Shaw, K. (2003). 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., & Meckling, W. (1976). Theory of the firm: Managerial behavior, agency costs and ownership structure. *Journal of Financial Economics, 3*(4), 305–360.
Jensen, N., Lyons, E., Chebelyon, E., Le Bras, R., & Gomes, C. (2020). Conspicuous monitoring and remote work. *Journal of Economic Behavior & Organization, 176*, 489–511.
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.
Larkin, I., & Leider, S. (2012). Incentive schemes, sorting, and behavioral biases of employees: Experimental evidence. *American Economic Journal: Microeconomics, 4*(2), 184–214.
Liang, C., Peng, J., Hong, Y., & Gu, B. (2022). The hidden costs and benefits of monitoring in the Gig economy. *Information Systems Research*. https://doi.org/10.1287/isre.2022.1130
Lord, F. M. (1967). A paradox in the interpretation of group comparisons. *Psychological Bulletin, 68*(5), 304.
Maas, V. S., & Van Rinsum, M. (2013). How control system design influences performance misreporting. *Journal of Accounting Research, 51*(5), 1159–1186.
Masella, P., Meier, S., & Zahn, P. (2014). Incentives and group identity. *Games and Economic Behavior, 86*, 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.
Pierce, L., Snow, D. C., & McAfee, A. (2015). Cleaning house: The impact of information technology monitoring on employee theft and productivity. *Management Science, 61*(10), 2299–2319.
Riener, G., & Wiederhold, S. (2016). Team building and hidden costs of control. *Journal of Economic Behavior & Organization, 123*, 1–18.
Schmelz, K. (2021). Enforcement may crowd out voluntary support for covid-19 policies, especially where trust in government is weak and in a liberal society. *Proceedings of the National Academy of Sciences, 118*(1), e2016385118.
Schmelz, K., & Bowles, S. (2021). Overcoming covid-19 vaccination resistance when alternative policies affect the dynamics of conformism, social norms, and crowding out. *Proceedings of the National Academy of Sciences, 118*(25), e2104912118.
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.
Snowberg, E., & Yariv, L. (2018). *Testing the waters: Behavior across participant pools* (Working Paper No. 24781). National Bureau of Economic Research.
StataCorp. (2019). *Stata statistical software: Release 16*. College Station, TX: StataCorp LLC.
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.

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