For online publication: Appendix

A Overview over the literature

Table A.1 provides a concise overview of the literature on the effects of control in experimental economics and illustrates our contribution.

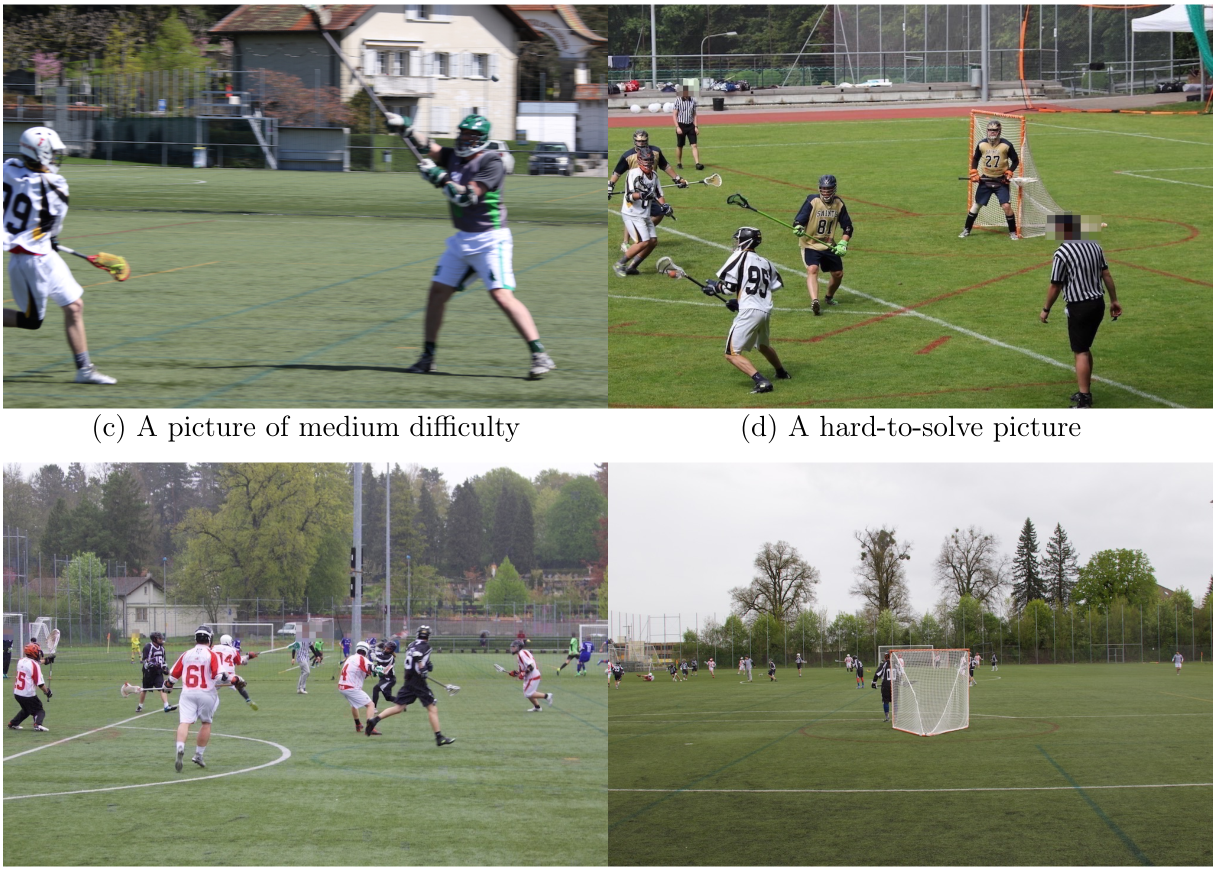
Table A.1: Overview of the literature in experimental economics

		Why do adverse effects of control arise?					Where do adverse effects of control arise?					
Article	Setting	Effectiveness of control technology	Legitimacy of control	Nature of P-A relationship	Procedural fairness	Reciprocity	Agents with positive attitude towards principal	Motivated agents (stated effort)	Motivated agents (real effort)	Non-controlled dimension	Hard tasks	Laborious tasks
Lab evidence:												
Falk and Kosfeld (2006)	Lab	✓						✓				
Dickinson and Villeval (2008)	Lab	✓		✓		✓			✓			
Schnedler and Vadovic (2011)	Lab		✓					✓				
Ziegelmeyer et al. (2012)	Lab	✓						✓				
Masella et al. (2014)	Lab			✓				✓				
Kessler and Leider (2016)	Lab				✓			✓				
Riener and Wiederhold (2016)	Lab	✓		✓				✓				
Burdin et al. (2018)	Lab					✓		✓				
Schmelz and Ziegelmeyer (2020)	Lab	✓		✓		✓		✓				
Field evidence:												
Nagin et al. (2002)	Field	✓					✓					
Boly (2011)	Field	✓										
Belot and Schröder (2016)	Framed Field	✓								✓		
This paper	Field								✓	✓	✓	✓

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

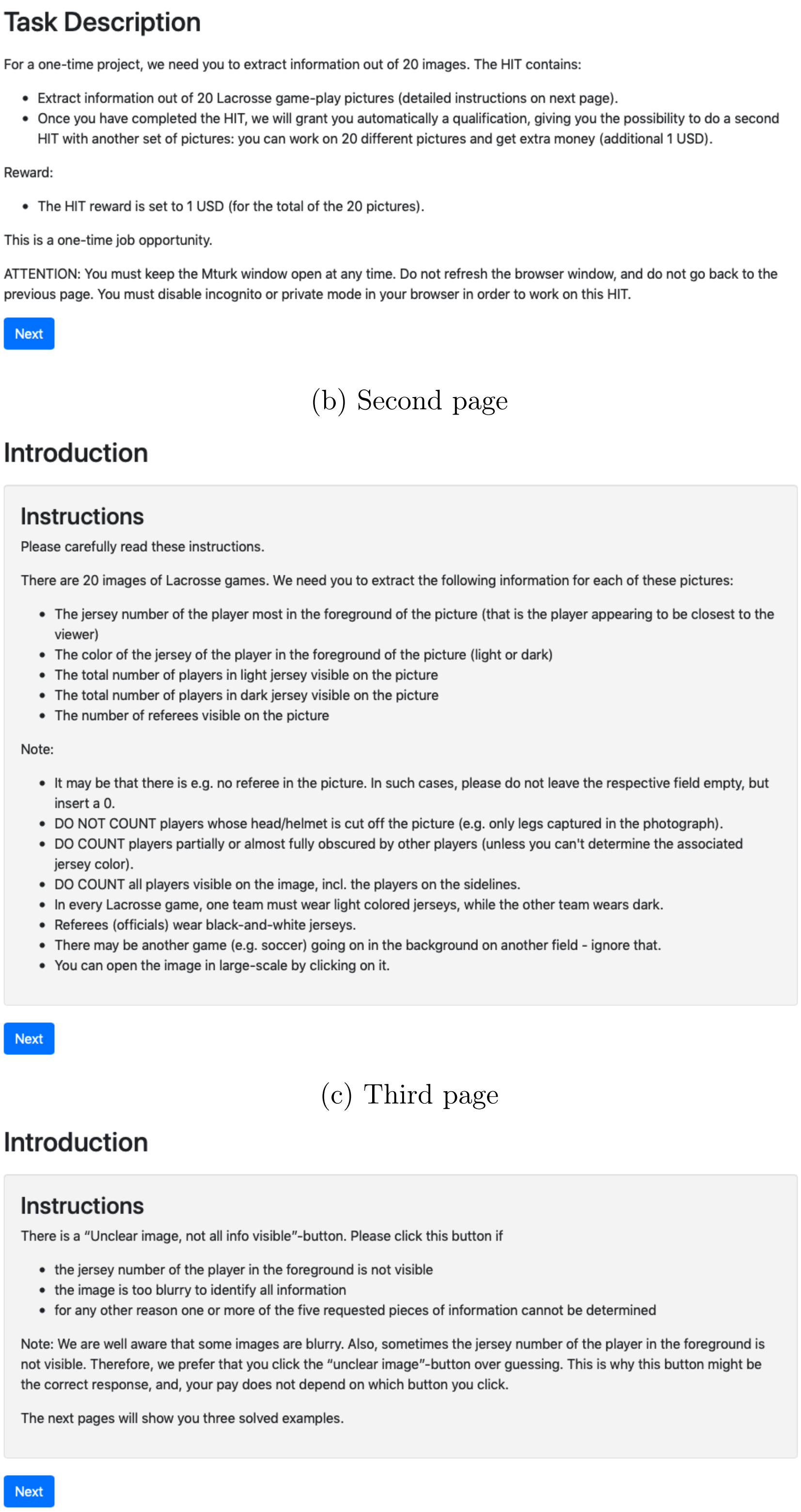
(d) A hard-to-solve picture

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.

(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.

(c) Third page

Introduction

Instructions

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

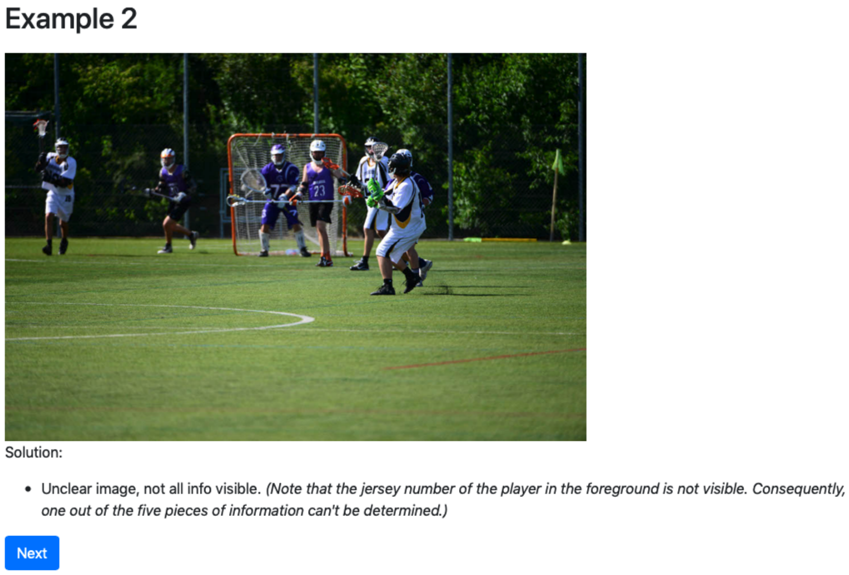
- the jersey number of the player in the foreground is not visible
- the image is too blurry to identify all information
- for any other reason one or more of the five requested pieces of information cannot be determined

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.

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

(a) Fourth page

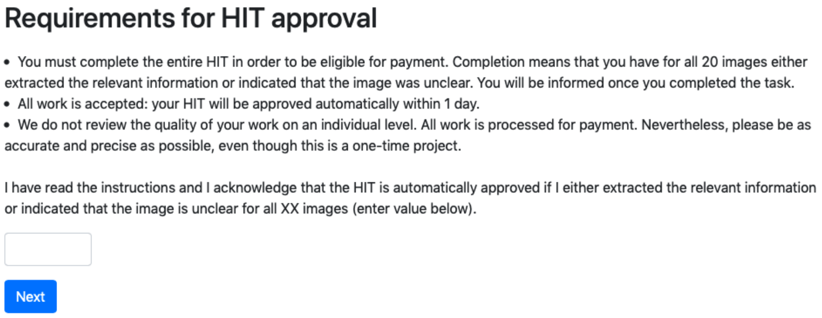


Example 2

Solution:

- Unclear image, not all info visible. *(Note that the jersey number of the player in the foreground is not visible. Consequently, one out of the five pieces of information can't be determined.)*

(b) Fifth page



Requirements for HIT approval

- You must complete the entire HIT in order to be eligible for payment. Completion means that you have for all 20 images either extracted the relevant information or indicated that the image was unclear. You will be informed once you completed the task.
- All work is accepted: your HIT will be approved automatically within 1 day.
- We do not review the quality of your work on an individual level. All work is processed for payment. Nevertheless, please be as accurate and precise as possible, even though this is a one-time project.

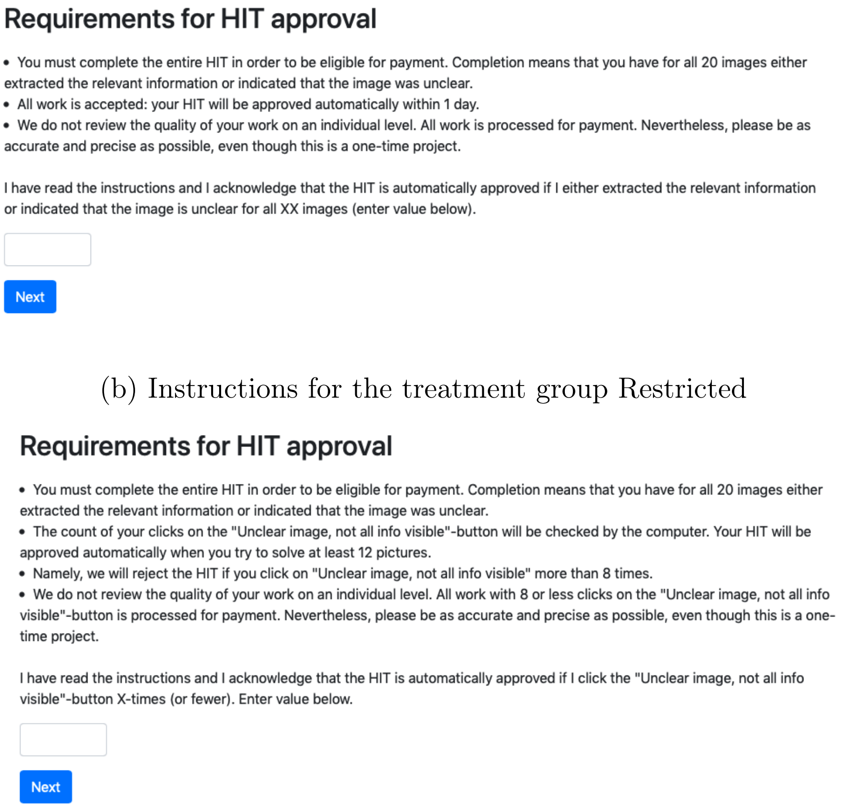
I have read the instructions and I acknowledge that the HIT is automatically approved if I either extracted the relevant information or indicated that the image is unclear for all XX images (enter value below).

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 Restricted group.

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

(a) Instructions for the Baseline group



Requirements for HIT approval

- You must complete the entire HIT in order to be eligible for payment. Completion means that you have for all 20 images either extracted the relevant information or indicated that the image was unclear.
- All work is accepted: your HIT will be approved automatically within 1 day.
- We do not review the quality of your work on an individual level. All work is processed for payment. Nevertheless, please be as accurate and precise as possible, even though this is a one-time project.

I have read the instructions and I acknowledge that the HIT is automatically approved if I either extracted the relevant information or indicated that the image is unclear for all XX images (enter value below).

(b) Instructions for the treatment group Restricted

Requirements for HIT approval

- You must complete the entire HIT in order to be eligible for payment. Completion means that you have for all 20 images either extracted the relevant information or indicated that the image was unclear.
- The count of your clicks on the "Unclear image, not all info visible"-button will be checked by the computer. Your HIT will be approved automatically when you try to solve at least 12 pictures.
- Namely, we will reject the HIT if you click on "Unclear image, not all info visible" more than 8 times.
- We do not review the quality of your work on an individual level. All work with 8 or less clicks on the "Unclear image, not all info visible"-button is processed for payment. Nevertheless, please be as accurate and precise as possible, even though this is a one-time project.

I have read the instructions and I acknowledge that the HIT is automatically approved if I click the "Unclear image, not all info visible"-button X-times (or fewer). Enter value below.

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

C Further Results

C.1 Descriptive statistics

Table C.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. Restricted 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 $p = .08$. As discussed in Section 2.3.3, we do not find any difference in attrition between the two treatment groups. We therefore conclude that the marginally significant pre-treatment differences are due to chance. Note that our analysis method, the difference-in-difference analyses and regressions conditional on pre-treatment measures, correct for this.

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 Restricted, and only 3.8% of restricted 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.

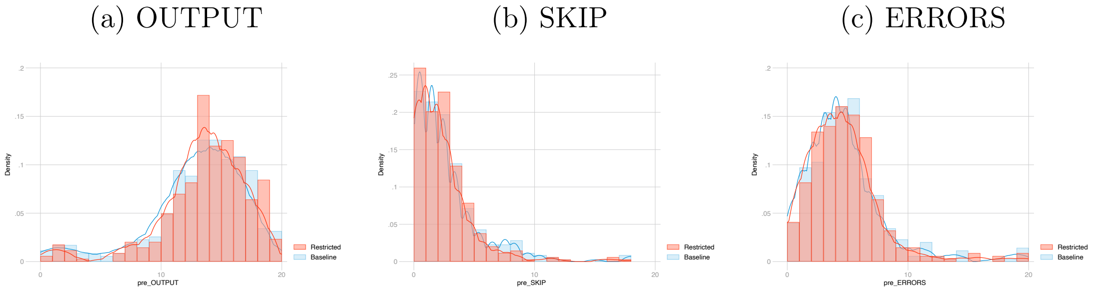
Table C.1: Descriptive statistics

	Pre-treatment stage		Experimental stage		Difference	
	Baseline	Restricted	Baseline	Restricted	Baseline	Restricted
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 Restricted 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 Restricted in the experimental stage.

Figure C.1 plots the distribution of the outcome variables in the pre-treatment stage, and Figure C.2 does so for the experimental stage.

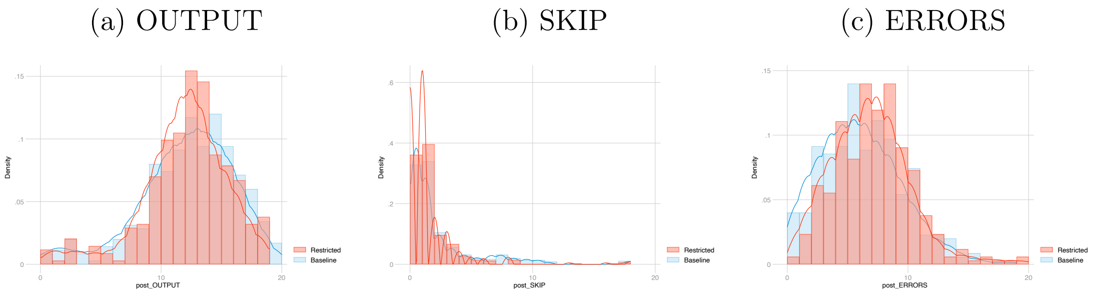
Figure C.1: Distribution of OUTPUT, SKIP and ERRORS in the pre-treatment stage


(a) OUTPUT

(b) SKIP

(c) ERRORS

Figure C.2: Distribution of OUTPUT, SKIP and ERRORS in the experimental stage


(a) OUTPUT

(b) SKIP

(c) ERRORS

C.2 Control Reduces Performance

Table C.2 test the robustness of our results reported in Figure 2 by regressing experimental stage measurements on the treatment dummy while conditioning on the pre-treatment stage measurements to control for individual pre-treatment characteristics.

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

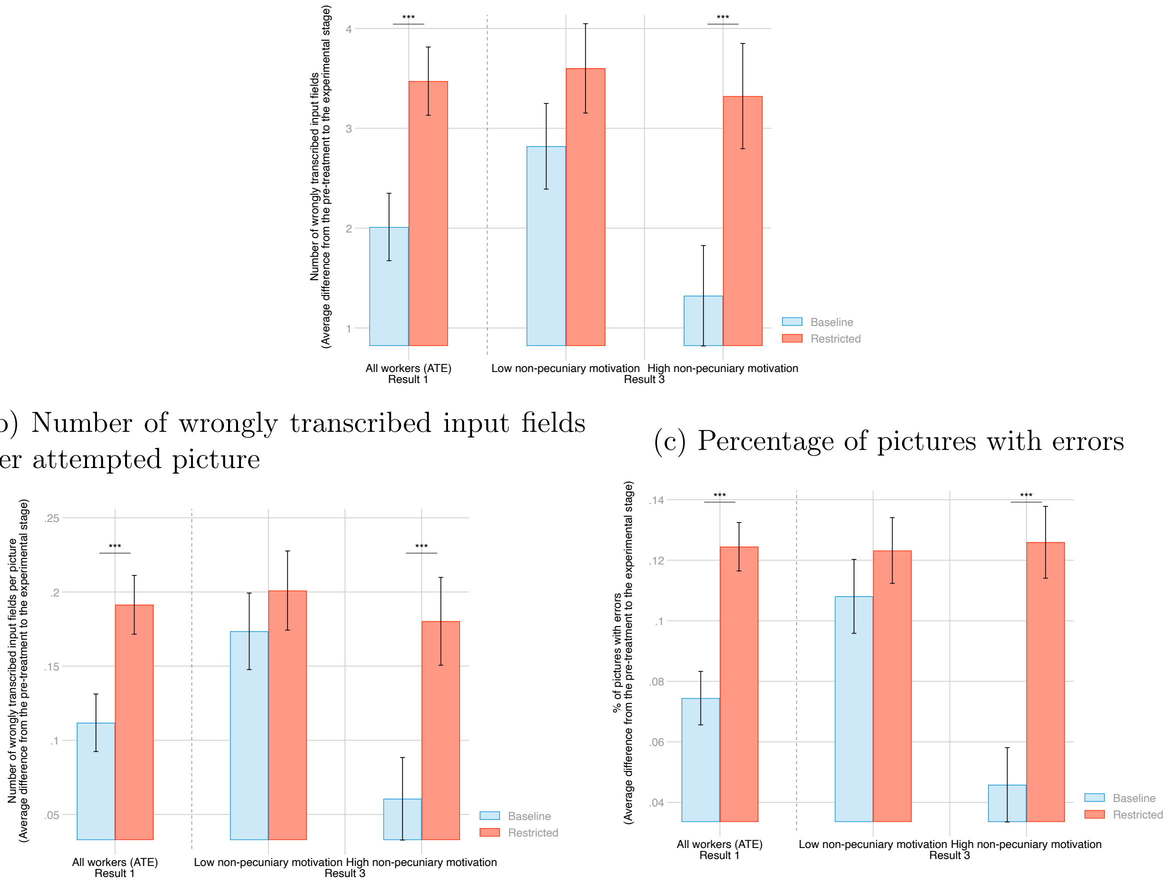
	(1) OUTPUT	(2) SKIP	(3) ERRORS
Restricted	-0.56	-0.38	0.92
	(0.18)	(0.13)	(0.19)
OUTPUT (pre-treatment)	0.74		
	(0.03)		
SKIP (pre-treatment)		0.74	
		(0.05)	
ERRORS (pre-treatment)			0.66
			(0.04)
Constant	2.49	0.14	2.89
	(0.42)	(0.13)	(0.23)
r2	0.59	0.56	0.42
N	693	693	693

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.

Figure 3 depicts the distribution of correctly solved pictures for the Baseline and the Restricted treatment in the experimental stage, and Figures C.1 and C.2 display the distribution plots of SKIP and ERRORS.

Figure C.3 employs alternative measures for ERRORS. Instead of a dichotomous classification of a picture as correct or false, Figure C.3a reports the average number of wrongly transcribed input fields (there are five input fields per picture), Figure C.3b reports the number of wrongly transcribed input fields per attempted picture (that is per non-skipped picture) and Figure C.3c 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 C.3: 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

C.3 Control Reduces Performance Among Workers With Non-pecuniary Motivation

Table C.3 continues with regression analysis to test the robustness of Result 2 reported in the main body, and regresses our outcome variables of interest on individual non-pecuniary motivation. Column (1) in Table C.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 Restricted workers: the coefficient of the interaction term between the Restricted group dummy and non-pecuniary motivation is negative and statistically

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

	(1)	(2)
	OUTPUT	
Restricted	0.85	0.03
	(0.49)	(0.26)
Non-pecuniary motivation, cont	0.18	
	(0.05)	
Restricted × Non-pecuniary motivation, cont	-0.20	
	(0.07)	
Non-pecuniary motivation (=1)		0.98
		(0.27)
Restricted × 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.

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 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.

C.3.1 Alternative Proxy Variables for Non-Pecuniary Motivation

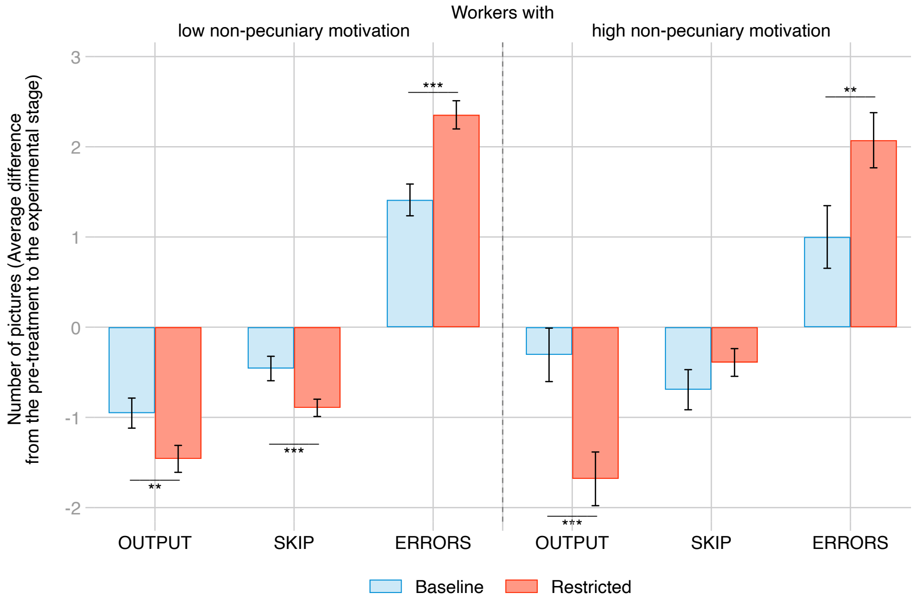
We test the robustness of Result 2 with alternative proxy variables for non-pecuniary motivation. Figure C.4 shows results when workers are classified into two types, those with high motivation and those with low motivation, based on whether they re-consulted in the pre-treatment stage the instructional 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 C.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.

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% ($p = .02$).

Regression table C.4 confirms this result. The interaction term of the Restricted treatment and non-pecuniary is negative, meaning that Restricted workers that were classified as non-pecuniary motivated decrease OUTPUT more than others ($p = .07$).

Figure C.5 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 C.5 plots

Figure C.4: 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, Restricted n=274. High non-pecuniary motivation N=144, whereof Baseline n=75, Restricted n=69.

Welch's t-test p values: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table C.4: Regression Analysis: Non-pecuniary motivation proxied by clicks on the "Open Instructions" -button

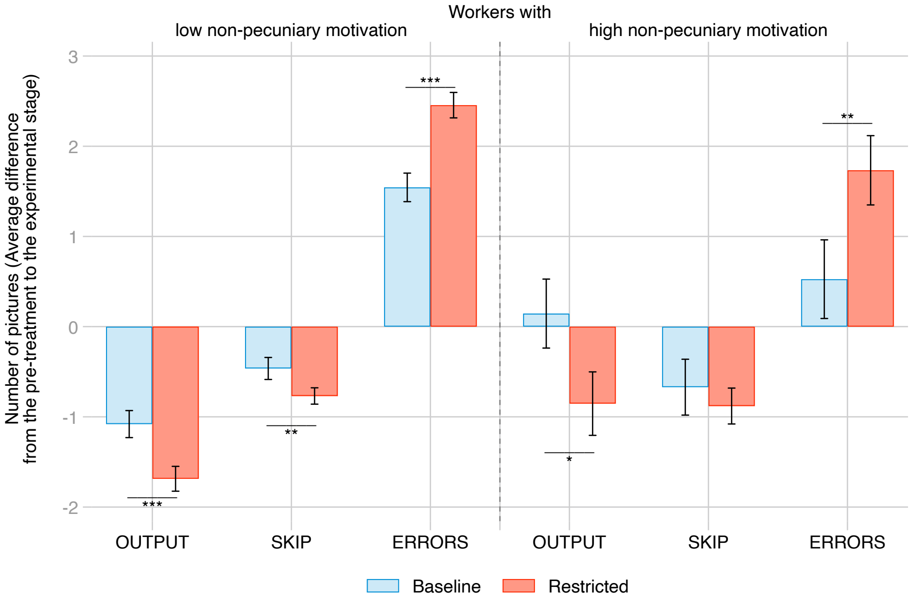
	(1) OUTPUT	(2) SKIP	(3) ERRORS
Restricted	-0.39	-0.54	0.93
	(0.21)	(0.16)	(0.22)
Non-pecuniary motivation (=1)	0.92	-0.43	-0.50
	(0.32)	(0.23)	(0.35)
Restricted × Non-pecuniary motivation (=1)	-0.74	0.77	-0.08
	(0.44)	(0.27)	(0.48)
OUTPUT (pre-treatment)	0.74		
	(0.03)		
SKIP (pre-treatment)		0.74	
		(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.

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 restricted by approx. 8.9% ($p = .06$). For workers with low motivation, depicted in the left panel, the performance decrease under control only amounts to approx. 4.7% ($p < .01$).

Figure C.5: 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, Restricted n=268. High non-pecuniary motivation N=151, whereof Baseline n=76, Restricted n=75.
Welch's t-test p values: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

In regression table C.5, we observe the interaction term of the Restricted treatment and non-pecuniary motivation to be negative. Again, this means that Restricted 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

Table C.5: Regression Analysis: Non-pecuniary motivation proxied by familiarity with the sport lacrosse

	(1) OUTPUT	(2) SKIP	(3) ERRORS
Restricted	-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)
Restricted × 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

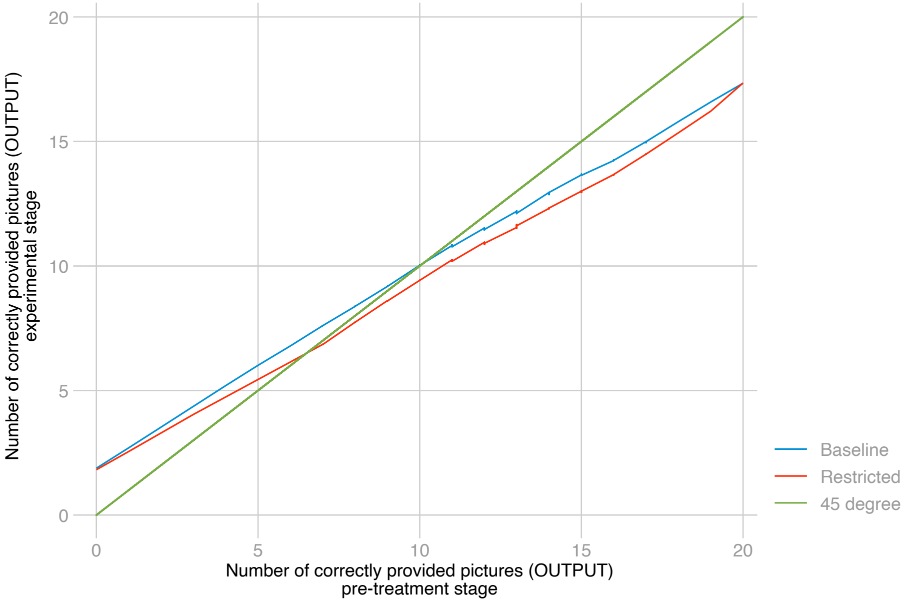
Note: OLS regressions, robust standard errors (in parentheses). Outcome variables are experimental stage measurements. OUTPUT: Number of correctly solved pictures. SKIP: Number of readable pictures that were declared as unreadable. ERRORS: Number of transcribed pictures that contain an error. A worker 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.

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 C.6 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 Restricted 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 C.6: Performance in pre-treatment stage vs. performance in experimental stage



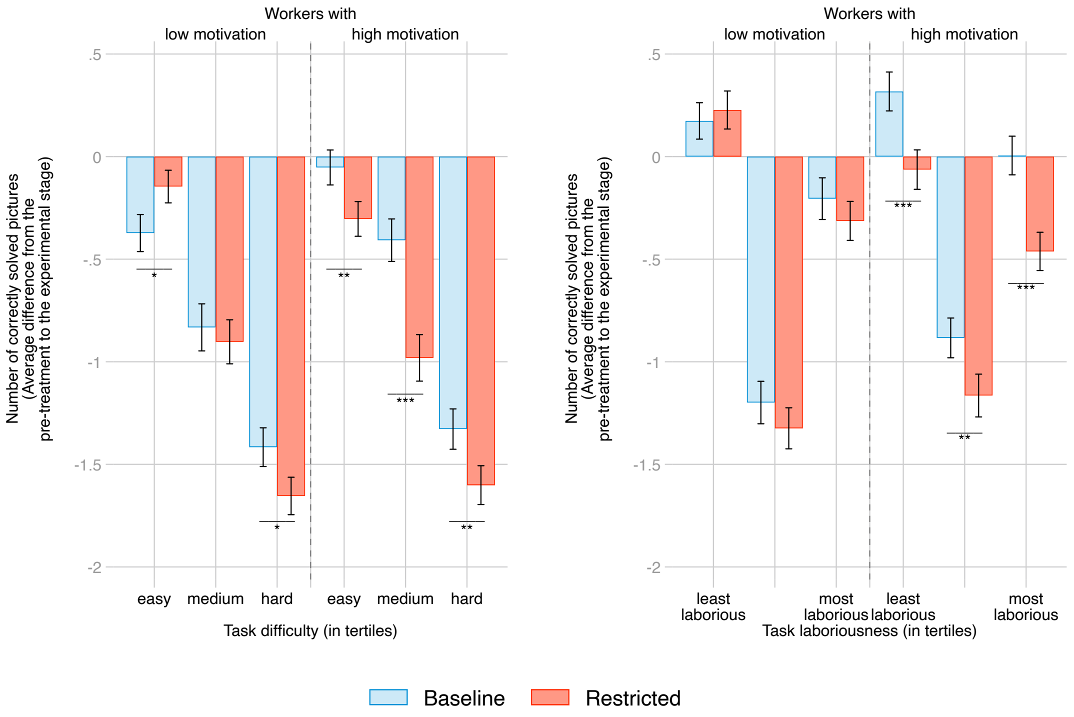
Note: By treatment group, the graph reports a locally weighted regression (default bandwidth) of performance in stage 2 against performance in stage 1. The graph reports on the vertical axis the number of correctly provided pictures in the experimental stage, and on the horizontal axis the number of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

C.3.2 Performance by task heterogeneity and type of worker

Let us revisit Result 3. 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 C.7), we find that Restricted 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, Restricted 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 ($p = .04$), to 0.57 pictures or 13.3% among medium pictures ($p < .01$) and to 0.27 pictures or 19.3% among challenging pictures ($p = .05$). Thus, Figure C.7 provides support that the performance reduction among hard and labor-intensive tasks is primarily driven by the motivated workforce.

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 among the motivated workforce is with 7.0% smallest among pictures that require the least effort ($p < .01$) and with 19.1%

Figure C.7: Performance by task heterogeneity and type of worker

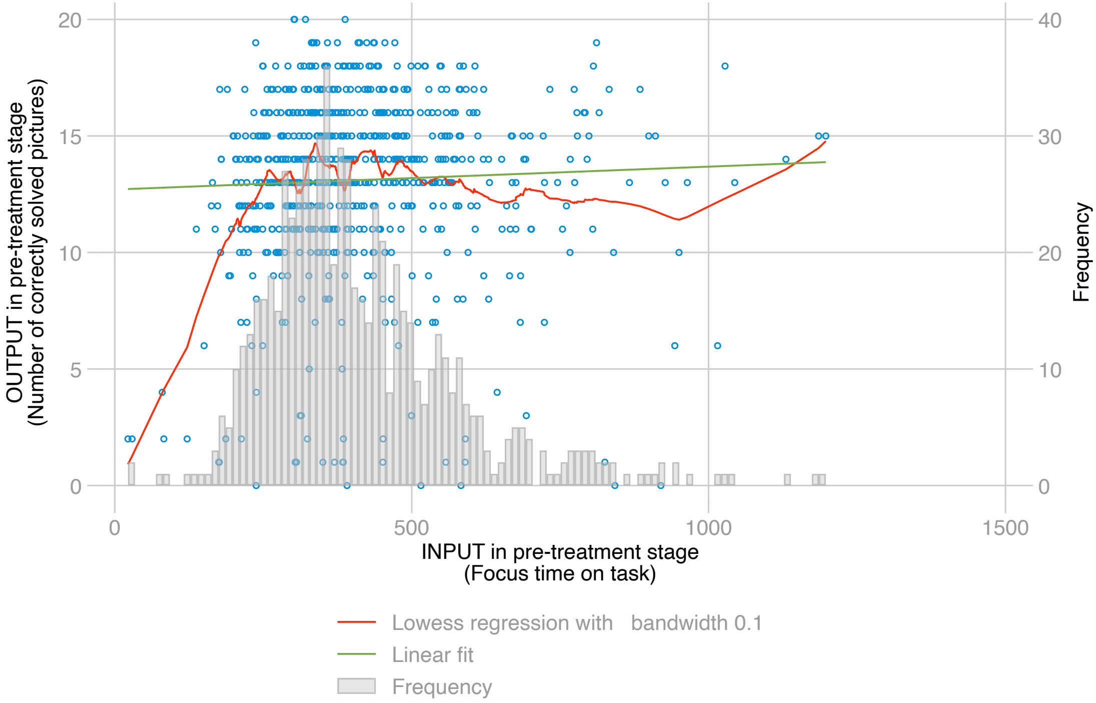


Note: The graph reports on the vertical axis the number of correctly transcribed pictures (OUTPUT) as an average difference from the pre-treatment to the experimental stage, representing the change in performance. The left panel reports the performance difference by task difficulty, the lower panel by 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, Restricted n=185. High non-pecuniary motivation N=347, whereof Baseline n=189, Restricted=158.

largest among the most time-demanding pictures ($p < .01$).

C.3.3 Relationship between performance and time on task (pre-treatment stage)

Figure C.8: 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.

C.4 Control Reduces Worker Performance Among Challenging Tasks

To assess the robustness of our Results 3 reported in the main body, we turn to regression analysis and estimate the models shown in Table C.6. 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), Restricted 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 ($p = .02$) 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 Restricted workers reduce performance by 0.17 pictures among the medium laborious category ($p = .06$) and substantially by 0.25 pictures among the labor-intensive tasks ($p < .01$).

Table C.6: 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
Restricted	-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.

D 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.

D.1 Results of Study 1 (original experiment)

Table D.1: Descriptive statistics, study 1

	Pre-treatment stage		Experimental stage		Difference	
	Baseline	Restricted	Baseline	Restricted	Baseline	Restricted
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					

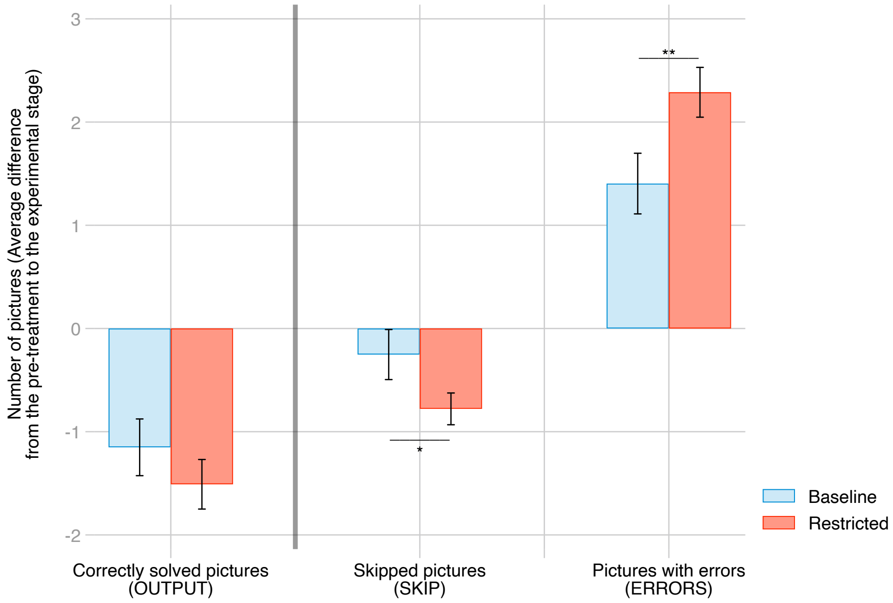
Note: For study 1, the table displays the means along with the associated standard deviation (in parentheses) for the pre-treatment stage, the experimental stage, and the difference between the two stages. Note that workers were randomized into Baseline and Restricted only in the experimental stage. 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 Restricted in the experimental stage.

D.1.1 Control Reduces Performance

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

Figure D.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 Restricted 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.

Figure D.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: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

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-restricted dimension. In the restricted dimension (number of skipped pictures), the control device has a small positive disciplining effect ($p = .07$). With regard to the non-restricted dimension, we observe a decline: The number of transcribed pictures that contain errors is significantly lower among restricted workers. Restricted workers submit on average 2.3 more pictures with transcription errors in the experimental stage, while non-restricted workers do so by 1.4 pictures only - a significant difference of 0.9 additional erroneously coded pictures ($p = .02$).

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

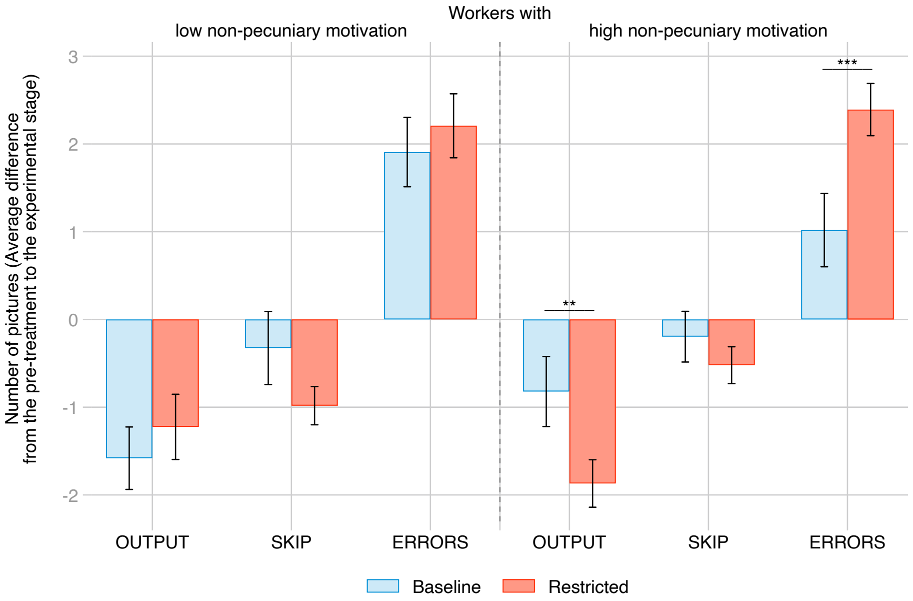
	(1)	(2)	(3)
	OUTPUT	SKIP	ERRORS
Restricted	-0.11	-0.65	0.75
	(0.35)	(0.27)	(0.36)
OUTPUT (pre-treatment)	0.64		
	(0.09)		
SKIP (pre-treatment)		0.66	
		(0.13)	
ERRORS (pre-treatment)			0.58
			(0.11)
Constant	3.72	0.45	3.26
	(1.24)	(0.27)	(0.53)
r2	0.38	0.31	0.24
N	203	203	203

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.

D.1.2 Control Reduces Performance Among Workers with Non-pecuniary Motivation

As formulated in Hypothesis 2, 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.

Figure D.2: 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 $n = 43$, Restricted $n = 58$. High non-pecuniary motivation $N = 102$, whereof Baseline $n = 56$, Restricted $n = 46$. Welch's t-test p values: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Support for Result 2 can be seen in Figure D.2 displaying the number of correctly solved pictures: Whereas motivated workers in the Baseline reduce their output by approximately 0.8 correctly solved pictures, motivated workers in the Restricted treatment reduce output by more than 1.9 correctly solved pictures, a significant difference of more than 1 picture ($p = .03$). For workers with low non-pecuniary motivation, we find no statistically significant differences. This two findings taken together, we find the negative effect of control on motivated workers to be significantly stronger than the negative effect of control on workers with low motivation ($p = .05$).

Figure D.2 also displays the number of readable pictures that were declared as unreadable. We do not observe a heterogeneous reaction in the Restricted dimension conditional on non-pecuniary motivation. When looking at the non-restricted task dimension, namely the number of pictures that were transcribed erroneously, we find that in the experimental stage, motivated workers in the Restricted 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 D.3 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.

D.1.3 Control Reduces Performance Among Challenging Tasks

Support for Result 3 is shown in Figure D.3, 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, Restricted workers perform worse than the Baseline by 0.32 pictures or 24.1% ($p = .06$).

The right panel in Figure D.3 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 Restricted workers is especially pronounced among pictures that require more labor. While the performance reduction of the Restricted 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 ($p = .09$).

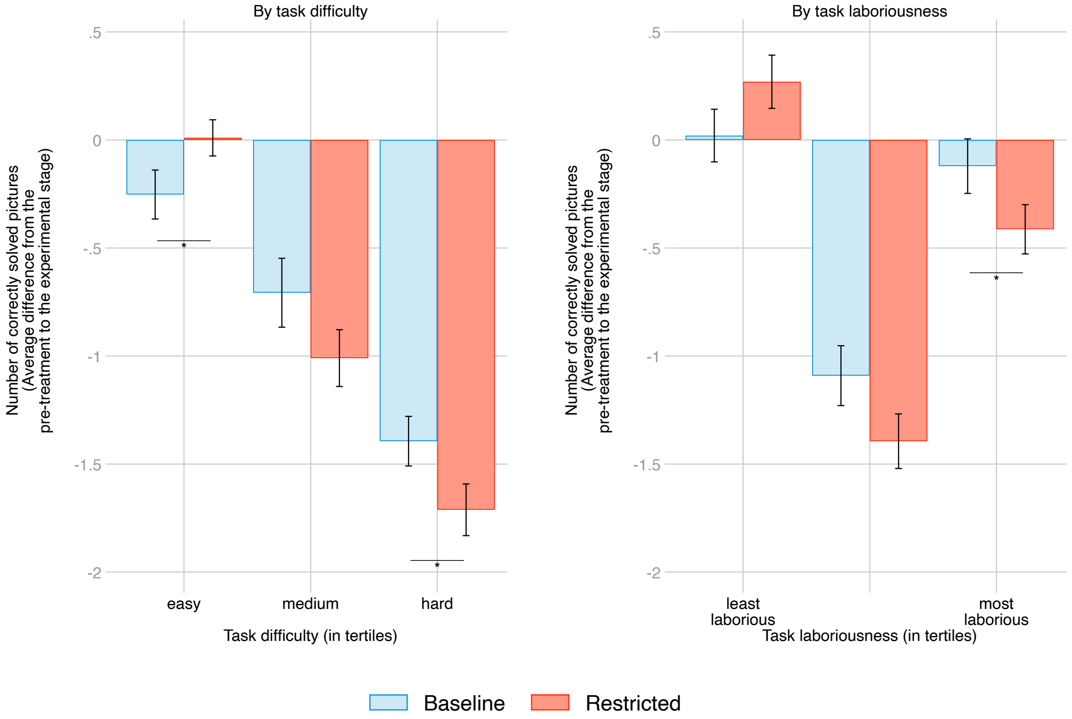
To asses the robustness of our results, we turn to regression analysis and estimate the

Table D.3: Regression Analysis: Non-pecuniary motivation interacted with treatment, study 1

	(1) OUTPUT	(2) SKIP	(3) ERRORS
Restricted	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)
Restricted × 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.

Figure D.3: 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 $n = 99$, Restricted $n = 104$.

models shown in Table D.4.

Table D.4: 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
Restricted	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), Restricted workers perform actually better than Baseline workers. The performance reduction occurs among the hard pictures (column (3)). This confirms Result 3: The control device reduces performance in the hard picture category by 0.28 pictures ($p = .07$), conditional on the pre-treatment performance. Again, similar results emerge when we order pictures according to task laboriousness. Restricted 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.

D.2 Results of study 2 (the repetition)

D.2.1 Control Reduces Performance

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

Figure D.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 Restricted group decrease the number of correctly solved

Table D.5: Descriptive statistics, study 2

	Pre-treatment stage		Experimental stage		Difference	
	Baseline	Restricted	Baseline	Restricted	Baseline	Restricted
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					

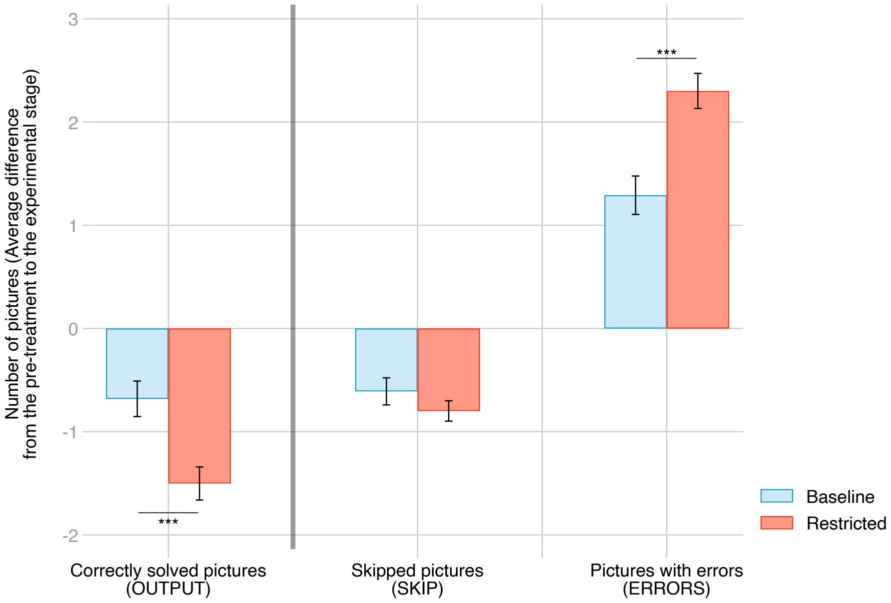
Note: For study 2, the table displays the means along with the associated standard deviation (in parentheses) for the pre-treatment stage, the experimental stage, and the difference between the two stages. Note that workers were randomized into Baseline and Restricted only in the experimental stage. 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 Restricted in the experimental stage.

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-restricted dimension. In the restricted dimension (number of skipped pictures), the restricted device has no significant effect. With regard to the non-restricted dimension, we observe that the number of transcribed pictures that contain errors is significantly higher among restricted workers: Restricted workers submit on average 2.3 more pictures with transcription errors in the experimental stage, while non-restricted 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 D.6 confirms these results.

Figure D.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 $n = 251$, Restricted $n = 239$.
Welch's t-test p values: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

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

	(1) OUTPUT	(2) SKIP	(3) ERRORS
Restricted	-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.

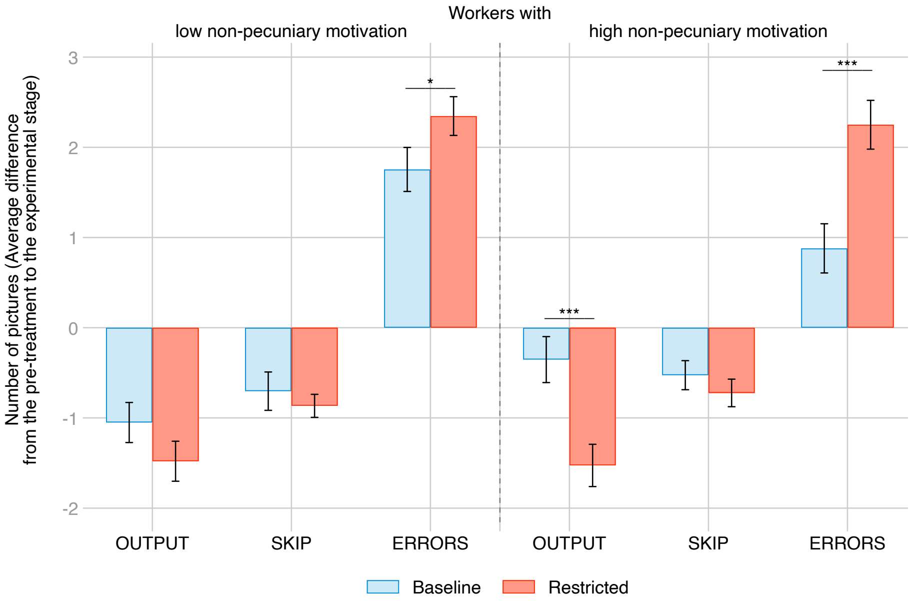
D.2.2 Control Reduces Performance Among Workers with Non-pecuniary Motivation

As formulated in hypothesis 2, 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 2.

Support for result 2 can be seen in Figure D.5 displaying the number of correctly solved pictures: Whereas motivated workers in the Baseline reduce their output by approximately 0.35 correctly solved pictures, motivated workers in the Restricted treatment reduce output by more than 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 (SKIP). We do not observe a heterogeneous reaction in the restricted dimension conditional on non-pecuniary motivation. The rightmost bars depict the non-restricted task dimension, namely the number of pictures that were transcribed erroneously: In the experimental stage, motivated workers in the Restricted treatment increase the number of pictures that contain errors by 2.3. Yet, motivated workers in the

Figure D.5: 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 $n = 118$, Restricted $n = 127$. High non-pecuniary motivation $N = 245$, whereof Baseline $n = 133$, Restricted $n = 112$. Welch's t-test p values: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

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-restricted 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 D.7. Column (1) reports regressions on the number of correctly solved pictures. It can be seen that the coefficient on the interaction term between the Restricted 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 D.7: Regression Analysis: Non-pecuniary motivation interacted with treatment, study 2

	(1) OUTPUT	(2) SKIP	(3) ERRORS
Restricted	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)
Restricted × 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.

D.2.3 Control Reduces Performance Among Challenging Tasks

We now turn to our third hypothesis, namely that the performance reduction particularly arises in more challenging tasks. Our findings are summarized in result 3.

Support for Result 3 is shown in Figure D.6, 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 Restricted workers solve 0.9 fewer pictures. Restricted 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. Restricted workers perform worse compared to the Baseline by 0.24 pictures, which represents a substantial performance reduction of 18.8% ($p = .04$).

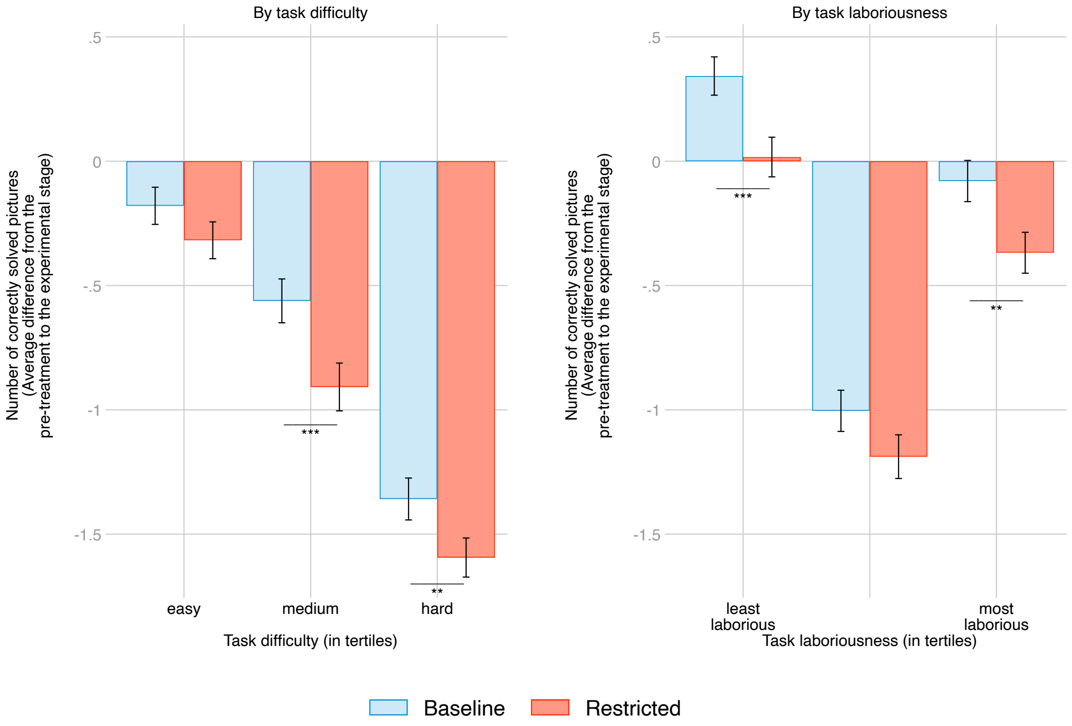
The right panel in Figure 5 plots a similar graph but by task laboriousness instead of task difficulty: Pictures are ordered into laboriousness tertiles based on the average time spent on a picture in the Baseline group. A similar pattern emerges. We observe that the performance reduction of Restricted workers is especially pronounced among pictures that require more labor. While the performance reduction of the Restricted group compared to the Baseline amounts to 0.33 pictures or 6.2% in the least laborious category ($p < .01$), it amounts to 0.29 pictures or 13% among the most labor-intensive pictures ($p = .02$).

To asses the robustness of our results, we turn to regression analysis and estimate the models shown in Table D.8.

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), Restricted 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: The control device reduces performance in the medium picture category by 0.29 pictures ($p = .02$) and in the hard picture category by 0.22 pictures ($p = .03$), 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. Restricted 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.

Figure D.6: 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 $n = 251$, Restricted $n = 239$.

Table D.8: 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
Restricted	-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.