

Asymmetric labor-supply responses to wage-rate changes.

Evidence from a field experiment*

PHILIPP DOERRENBERG DENVIL DUNCAN MAX LÖFFLER

This version: November 2015

First version: August 2015

Abstract

The standard labor-supply literature typically assumes that the labor supply response to wage increases is the same as that for equivalent wage decreases. However, evidence from the behavioral-economics literature suggests that people are loss averse and thus perceive losses differently than gains. This behavioral insight may imply that workers respond differently to wage increases than to wage decreases. We estimate the effect of wage increases and decreases on labor supply using a randomized field experiment with workers on Amazon's Mechanical Turk. The results provide evidence that wage increases have smaller effects than wage decreases, suggesting that the labor-supply response to wage changes is asymmetric. This finding is especially strong on the extensive margin where the elasticity for a wage decrease is twice that for a wage increase. These findings suggest that a reference-dependent utility function that incorporates loss aversion is the most appropriate way to model labor supply.

JEL Classification: J22, J31, D01

Keywords: Labor supply, loss aversion, labor supply elasticities w.r.t. wages

***Doerrenberg:** ZEW Mannheim and IZA Bonn. doerrenberg@zew.de. **Duncan:** Indiana University, Bloomington, ZEW and IZA. duncande@indiana.edu. **Löffler:** ZEW Mannheim and University of Cologne. loeffler@zew.de. David Vjazner provided excellent research assistance. We are grateful to Harald Lang, Andreas Peichl, Jan Schmitz, Sebastian Sieglöcher, Christopher Zeppenfeld, and seminar/conference participants at SPEA/IU (Bloomington), IIPF 2015 (Dublin) and ZEW Mannheim for valuable comments and suggestions. The usual disclaimer applies.

1 Introduction

Motivation and Research Question. The standard labor supply literature typically makes the implicit assumption that the labor supply response to wage increases is the same as that for equivalent wage decreases. In other words, wage increases and wage decreases of equal magnitude have the same effect (though with opposite signs) on labor supply decisions. This assumption implies that labor-supply elasticities with respect to wages do not depend on the sign of the wage variation. However, an extensive literature building on Kahneman and Tversky (1979) has established that individuals are loss averse and thus perceive negative changes (changes in the loss domain) differently than positive changes (changes in the gain domain). This behavioral insight suggests that workers respond differently to wage increases than to wage decreases, and thus casts doubts on the accuracy of labor supply elasticity estimates that do not account for the sign of the wage variation.

Although there are a few studies that incorporate loss aversion into empirical strategies aimed at identifying labor supply elasticities (related literature discussed below), there still remains a large gap in the literature regarding the symmetry of the wage elasticity of labor supply. This paper contributes to filling this gap in the literature. Our precise research question is: do wage increases and decreases of equal magnitude have symmetric effects on labor supply? Answering this research question requires a set-up where reforms introduce (quasi-) randomly assigned wage increases and decreases at the same time for comparable individuals. Finding such types of reforms in ‘natural’ settings is difficult, if not impossible, and thus may partly explain the sparse literature on the symmetry of labor supply responses to wages.

The Field Experiment. We address these empirical challenges using a field experiment on labor supply where we randomly assign wage increases and decreases of equal magnitude to workers. Specifically, we set-up a real labor task and invite workers to work on this task in an actual online-labor-market, namely Amazon’s Mechanical Turk (henceforth mTurk). The labor task is advertised on the mTurk website as any other labor task and workers receive wages that are comparable to other wages on mTurk. In addition, the labor task is designed to be perceived as realistic as possible; it requires workers to transcribe scanned German-language documents. Importantly, the workers in our experiment do not know that they are participating in an academic experiment.

The experiment is designed to induce an exogenously determined reference wage using two strategies. First, we announce a certain wage per transcribed picture in the advertisement of our task on the mTurk website. Second, workers complete one batch of transcriptions for the wage announced in the mTurk advertisement. After transcribing the first batch of images, all workers are randomly assigned to one of three groups: 1) the wage remains constant (control group), 2) the wage increases by 20%, 3) the wage

decreases by 20%. After the updated wages have been presented to workers, they can select to either stop working on our labor task or keep working as much as they wish. We identify the symmetry of the labor supply response by comparing labor supply behavior between the three randomly assigned groups.

Findings. The results show that the labor-supply response to wage changes is asymmetric. This asymmetry is especially strong on the extensive margin defined as the share of workers who quit conditional on seeing the treatment information. The estimated extensive-margin treatment effect for workers who experience a wage decrease is approximately twice that of workers who experienced a wage increase. Estimates of the intensive margin response are also suggestive of an asymmetric response; differences are large, but imprecisely estimated. Though we cannot reject the null that the number of transcriptions responded symmetrically to the wage change, we find that the share of the transcription response that can be explained by the extensive margin differs significantly between the wage increase and decrease groups. Finally, the results show that the wage changes did not have any effect on the quality of transcriptions, which is above 96 percent in all groups. Over all, our results point to an upward sloping supply curve that appears to be kinked around the reference wage of \$0.15.

Our findings can be rationalized by the theoretical model put forward by Ahrens et al. (2014). They incorporate loss aversion in a model of labor supply and show that the labor supply curve is kinked at a reference wage. Drawing on the loss aversion literature, they further show that the supply curve is flatter in the loss domain than in the gain domain. This implies that wage decreases have stronger labor-supply effects than wage increases; i.e., the labor supply response to wage changes is asymmetric even for marginal changes in wage. This is precisely what we find.

Contribution to the Literature. Our paper speaks and contributes to three strands of literature. First, we add to the literature on labor-supply effects of wage changes. Economists have explored the effect of wages on labor supply for several decades (see Keane, 2011 for a survey). Many of these studies use panel-data sets and exploit positive and negative variation in wages to estimate the wage elasticity of labor supply.¹ Because the elasticity estimated by these studies represents roughly an average of wage-increase-induced and wage-decrease-induced elasticities, our paper suggests that existing estimates likely overestimate the effect of wage increases while underestimating the effect

¹It is sometimes argued that nominal wage cuts are rare and therefore not relevant. While we acknowledge that nominal wage cuts occur less often than increases, it has been shown that wage cuts do happen; for example during recessions and bankruptcies, and for the self-employed and salary earners (Kahn, 1997). In addition, many studies on labor-supply elasticities use upward and downward variation in tax rates to instrument for wages (e.g., Eissa and Liebman, 1996; Rothstein, 2010). This generates downward variation in wages even in the absence of nominal wage cuts. Our study is also relevant for decreases in real wages, which occur more frequently than nominal wage cuts. Our results may suggest that inflation-induced decreases of real wages have larger labor supply effects than previously thought.

of wage decreases.

Relatedly, our results further raise questions about the comparability of labor-supply elasticities across studies that differ in the sign of the wage changes used for identification. Our findings suggest that it cannot be concluded from the estimated elasticities that workers are more responsive in the one setting relative to another without knowing whether the sign of the wage changes is the same. This is especially important for meta-analysis studies on labor supply (e.g., Evers et al., 2008). Our findings imply that in such meta-analyses one should include an indicator variable to distinguish between labor supply estimates based on wage increases from those based on wage decreases. Our finding that the largest asymmetry is along the extensive margin is especially important for understanding the labor-supply effects of wages since it is generally accepted that labor-supply elasticities are mainly determined by the extensive margin response (Blundell and MaCurdy, 1999; Meghir and Phillips, 2010; Bargain et al., 2014).

The study most closely related to ours is Kube et al. (2013), who conduct a field experiment with students working in a library for a given period of time. They generate an exogenous reference wage by announcing a projected hourly wage to all workers when the job is advertised. Immediately before the task starts, they announce a higher wage to workers in one treatment group and a lower wage to workers in another group. Workers in the control condition receive the initially announced wage. The study finds that the wage cut decreases work effort (i.e., output generated during the given period of time) whereas the wage increase does not have any effect relative to the control group. In line with our findings on transcription accuracy, their study also does not find any effects on quality of work. While these results are broadly consistent with our findings, our paper differs from theirs in the design of the labor market institution; this has important implications for the interpretation and application of our findings. The institutions differ in that we pay workers for each transcribed picture instead of for a predetermined number of hours, and we allow workers to quit the labor task whenever they choose to do so. Furthermore, our analysis is based on a much larger sample of workers from a real-world labor market. Therefore, our design is representative of labor markets where workers receive piece-rate payment and have tremendous labor supply flexibility whereas Kube et al. (2013) focuses on labor markets where workers are required to work a predetermined number of hours for a fixed hourly wage rate. One advantage of our design is that workers are able to respond on two additional margins that are not included in Kube et al. (2013); extensive margin and the intensive-time margin. As a result, we are able to study asymmetric responses to wage changes on the extensive and intensive margins. Additionally, because our workers receive a piece rate, subjects who reduce output earn a lower pay-off and have less scope to punish their employer through shirking. This reduces the likelihood that our findings are driven by reciprocity as in Kube et al. (2013). Therefore, we are

able to show that labor supply asymmetry exists even in the absence of a reciprocity motive. The institutional frame-work of our study – large sample of workers in their natural labor-market environment – also implies that our findings can be generalized to similarly situated labor markets; large crowd-sourcing labor markets characterized by low wage and high flexibility.

Our paper further relates to several studies showing that individual labor supply decisions are affected by target incomes. In a survey of the literature, Goette et al. (2004) show how empirical results on labor-supply behavior are consistent with reference-dependent preferences where workers provide high effort if they are below a target income, whereas they provide less effort if they have surpassed a target. These types of preferences are, for example, found for taxi drivers (Camerer et al., 1997; Crawford and Meng, 2011) or bicycle messengers (Fehr and Goette, 2007). While these studies demonstrate that workers have target incomes, they do not allow conclusions about the asymmetric effects of wages.

Second, our paper makes a direct contribution to the behavioral-economics literature on loss aversion following Kahneman and Tversky (1979). This literature pursues the idea that individuals evaluate outcomes relative to reference points. These types of preferences are commonly termed reference-dependent preferences and have been formalized by Koszegi and Rabin (2006, 2007, 2009). Loss aversion, “the most notable manifestation of such reference-dependent preferences” (Koszegi and Rabin, 2006, page 1133), describes the notion that individuals weight negative deviations (losses) from the reference point more than gains of equal magnitude. In models of reference-dependent preferences, the reference point is usually assumed to be determined by the individual’s expectations. There is a large empirical literature showing that individuals indeed have preferences consistent with loss aversion and that individual expectations determine the reference point (e.g., Dunn, 1996; Post et al., 2008; Abeler et al., 2011; Card and Dahl, 2011; Marzilli Ericson and Fuster, 2011; Pope and Schweitzer, 2011). We add to this literature in that we provide additional empirical evidence that individuals have preferences that are consistent with loss aversion and reference dependence in the context of labor supply.

Finally, our paper raises important questions about the elasticity of taxable income which plays a crucial role in our understanding of the efficiency costs of taxation (e.g., Saez et al., 2012; Kleven and Schultz, 2014). In particular, our results suggest that failure to distinguish between ETI estimated with tax rate increases and ETI estimated with tax rate decreases is likely to lead to an underestimation of the efficiency cost of tax rate increases. This problem is likely to be even more important than with wage changes since tax rates generally move freely in both directions. Of course, the labor supply response to wage changes is not necessarily identical to the response to tax rate changes. Therefore, we are cautious in generalizing our results to the case of tax rate changes.

Structure of the Paper. The paper is organized as follows. Section 2 describes the real labor task and its implementation in Amazon’s Mechanical Turk. In Section 3 we lay out the theoretical approach following Ahrens et al. (2014), which incorporates loss aversion. Section 4 describes the data and our empirical approach. We present the results in Section 5 before 7 concludes.

2 The Experiment

This section describes the field experiment used to estimate the impact of wage rate changes on labor supply. We begin by describing the labor task, the treatment design and then implementation in Amazon’s Mechanical Turk.

2.1 Design

Labor Task. We selected an online labor task that requires subjects to transcribe German text shown a series of images. The German texts are taken from a recent publication, but each page of the document is deliberately ruffled so that the scanned versions appear much older than they really are. The advantage of changing the appearance of the images is the subjects are more likely to believe that the texts were scanned from old books for which a digital copy is not available. The task then, is to digitize these “old” German books. Each image has approximately five lines and 43 words (344 characters). Figure 1 shows an example. Subjects are randomly shown one of 128 images at a time and are instructed to hit “save picture” when they are done transcribing the text in the image. A new image is shown after the subject hits “save picture”.

Treatment Groups. We use a between-subjects design in order to identify the effect of wage changes on labor supply. Subjects are randomly assigned to one of three groups: one control group and two treatment groups. Subjects in all three groups work on the labor task described above and are paid a piece rate for each image that is transcribed. The piece rate (called *bonus* in the experiment) is set at \$0.15 for each of the first six transcribed images in all three groups. Subjects receive a notification thanking them for transcribing the images after the first six images have been transcribed. They are then told that they can transcribe additional images and that the piece rate for the additional images is either \$0.18, \$0.15 or \$0.12, for the wage-increase, control, and wage-decrease groups, respectively (see Figure 4 for an explanatory treatment notification). Notice that the wage rate remains fixed at \$0.15 for the control group, and that the wage rate change is the same for both treatment groups; in each case the rate changes by \$0.03 or 20%.

Reference Point. The experiment is designed to exogenously establish a clear and salient reference wage. The literature typically finds that reference points depend on rational, individual expectations, suggesting that expectations about the per-unit wage form the reference point in the context of labor supply decisions in our experiment (e.g., Koszegi and Rabin, 2006; Abeler et al., 2011; Ahrens et al., 2014). Therefore, potential workers are told that the wage per transcribed picture is \$0.15 in the job announcement. In addition, workers who start working on our task face the announced wage of \$0.15 for the first six transcribed pictures, after which the wage rate either increases or decreases. We argue that this design generates the expectation that the per-unit wage will remain constant at \$0.15 throughout the entire task. In other words, we argue that \$0.15 constitutes the reference point in our experimental set-up.

One potential drawback of our experimental design is that it may raise concerns of deception since the job description does not notify subjects of the possibility that the wage may increase or decrease after a certain number of transcribed pictures. This was a deliberate choice in an effort to establish a clear and salient reference point.² We minimize these deception concerns by including the following pieces of information in the treatment notification (see Figure 4). First, we thank the workers for completing the transcription task and remind them that, as promised in the introduction of the task, they will be paid \$0.15 for each of the pictures they transcribed so far. Next, we inform them that they have the option to transcribe additional images and that the piece rate for these additional transcriptions is different from that for the first batch of transcriptions. Finally, we make it clear that they can stop and exit the task at this point if they wish and instruct them on what to do next to ensure we are able to process their payment.³ We argue that these design features make it clear to workers that they first transcribe pictures based on the piece rate announced in the introduction to the task, and that they can transcribe additional pictures at a new rate.

2.2 Implementation

Labor Market and Recruitment. The experiment is implemented in the field using workers on Amazon’s Mechanical Turk. mTurk is an online labor market where job offers are posted and workers choose jobs for payment. It has numerous benefits for running

²If we had informed subjects about the possibility of a wage change, we would have generated uncertainty about the eventual wage and the reference wage would not have been as clear.

³The notification reads: “Thank you for transcribing these pictures. As written in the introduction, we will grant a bonus of \$0.15 for each of these pictures. There are additional pictures that you can transcribe. However, the bonus payment for each additional picture will be \$0.12/\$0.18 from now on. You will receive \$0.15 bonus for each of the six pictures you transcribed so far, though. If you want to stop and exit, just copy your Personal ID to the Amazon Turk Website and submit the HIT.” Instead of the wage change, we include the following message for the control group: “There are additional pictures that you can transcribe. Just as before, the bonus for each additional picture will be \$0.15.”

experiments, including access to a large stable subject pool, diverse subject background, and low cost.⁴ Furthermore, the behavior of online workers has been shown to be comparable to those of subjects in laboratory studies (Horton et al., 2011). In addition, experimenter effects are avoided because subjects do not know that they participate in an experiment (Paolacci et al., 2010; Horton et al., 2011; Buhrmester et al., 2011; Mason and Suri, 2011). Importantly for us, we are able to identify the effect of wages changes in a naturally occurring labor market. In general, experiments on Amazon Turk therefore combine internal and external validity since it is a real labor market with actual workers where randomized trials can be conducted (Horton et al., 2011).⁵

Although we recruit workers through mTurk, they complete the labor task on an external website that we created for the purposes of the experiment. We first create a human intelligence task (HIT) that is advertised on mTurk. The HIT includes a description of the labor task and compensation. It also includes instructions for how to complete the task; see Figure 2. Particularly, subjects are told to accept the HIT and click on the weblink if they are interested in completing the task. Subjects who click on the link are taken to our external website where they are randomly assigned to one of three groups and shown the instructions in Figure 3. Subjects are instructed to click continue if they wish to work on the task, and those who do are shown images of scanned German text that they must transcribe for payment. Each page of our website shows the subjects their personal ID, number of pictures transcribed so far, and the current piece rate. We implement treatment after six images have been transcribed and limit the total number of images that each subject can transcribe to 50. However, subjects are not aware of either of these limits until they reach them. In other words, subjects do not know that the HIT has six images, that they will have the opportunity to continue working after the first six images, that the piece rate might be different if they continue working, or that they can only transcribe up to 50 images if they chose to continue working. Subjects in wage-decrease group who complete six transcriptions are shown the treatment information illustrated in Figure 4. A similar text is shown to subjects in the wage-increase group and the control group; the only difference is the piece rate for the additional images.

Transcribing text from an image can be a tedious task. However, given that the text in the images is short, the task could be perceived as mostly costless for German speakers. In order to reduce this possibility and ensure that the labor costs are non-zero, we restrict the subject pool to workers with a US IP address. The idea here is that the labor cost of transcribing German text is much higher for non-Germans than for Germans. Of course, our restriction does not preclude the possibility that German speakers participated in the task. However, any Germans who participated in our experiment are randomly distributed

⁴According to Amazon, there are over 500,000 workers from 190 countries in the mTurk labor market: <https://requester.mturk.com/tour>.

⁵Kuziemko et al. (2015) is a recent example of an economics paper using Amazon's Mechanical Turk.

across our treatments and therefore have no effect on our outcome of interest.

The experiment is programmed on mTurk to expire after 750 workers accept the HIT or 10 days have passed, whichever comes first. Our initial run of the experiment, which started on June 15, 2015, expired after 10 days with only 418 workers. Therefore, we initiated a second run on July 20, 2015, and this run expired after hitting the 750 worker threshold six days later. In total, 1,168 workers participated in the two runs.

Payment. The experiment ends for each subject when she decides to stop or when she transcribes 50 pictures, whichever comes first. In either case, each subject is instructed to copy her personal ID number, which is shown in the top right corner of each page, and paste it in the entry box on the mTurk website. This process is necessary for us to match subjects to their mTurk worker ID and thus process their payments. Subjects receive a participation reward of \$0.10, which is paid as long as a subject accepts the HIT and completes at least one transcription. Additionally, subjects are paid a piece rate of \$0.15 for each of the first six transcribed pictures, and depending on treatment group, \$0.12, \$0.15 or \$0.18 for each transcribed image above the first six transcriptions. Given the payment restrictions imposed by the mTurk platform, we frame the piece rate as a bonus in all communications to the subjects. For example, subjects in the control group are told they will be paid \$0.10 for participating in our HIT and a bonus of \$0.15 for each transcribed picture.

3 Theoretical Framework

This section presents a theoretical framework that allows us to predict the impact of wage increases and decreases on labor-supply. The framework is informed by Ahrens et al. (2014) who incorporate loss aversion into a standard labor-supply model.

The Model. Ahrens et al. (2014) develop a model where workers with reference-dependent preferences maximize the following utility function:

$$U(C, L) = U^C(C) - \theta_i \frac{L^{\vartheta_i}}{\vartheta_i},$$

where C is consumption, L is labor supply (hours worked or effort), and θ_i is a parameter to ensure preference continuity at the reference wage. $U^C(C)$ is utility from consumption and the term $\frac{L^{\vartheta_i}}{\vartheta_i}$ indicates disutility from working. ϑ_i is a measure of loss aversion, which is characterized by the following piece-wise function:

$$\vartheta_i = \begin{cases} \vartheta_1 & \text{if } w > w^r \\ \vartheta_2 & \text{if } w < w^r. \end{cases}$$

In this equation, w is the current wage (per unit of L supplied) and w^r is the reference wage. If w is above the reference wage, the worker is in the so-called gain domain, and if w is below the reference wage, she is in the loss domain. A subject is loss averse if $\vartheta_1 > \vartheta_2$, implying that the marginal utility loss from working is higher in the gain domain than in the loss domain. This means that workers are less willing to supply an additional unit of labor when the wage is above the reference wage than when it is below. Maximizing with respect to the budget constraint $C = wL$ gives the following kinked labor-supply curve:⁶

$$L = \begin{cases} \left(\frac{w}{\theta_1}\right)^{\frac{1}{\vartheta_1-1}} & \text{if } w > w^r \\ \left(\frac{w}{\theta_2}\right)^{\frac{1}{\vartheta_2-1}} & \text{if } w < w^r \end{cases}$$

The Prediction. Because of loss aversion with respect to the reference wage w^r (and hence $\vartheta_1 > \vartheta_2$), we get that $\frac{1}{\vartheta_1-1} < \frac{1}{\vartheta_2-1}$. This implies that subjects whose current wage is the reference wage w^r are more responsive to wage decreases than to wage increases.⁷

The main insight from this theoretical framework is sketched in Figure 5, which relates leisure and wages. A worker who is located at the reference wage, denoted R , will respond differently to wage increases and decreases of equal magnitude. In particular, a worker at the reference point weights wage decreases more heavily than wage increases. As a result, she will respond more strongly to a wage decrease (by working less) than an equally sized wage increase (to which she will respond through more labor supply). This result implies that labor supply elasticities identified from wage increases are predicted to be smaller than labor supply elasticities identified from wage decreases. Our field experiment tests this prediction; the results are presented in the next sections.

The Reference Wage. The natural question at this point is regarding the determination of the reference wage w^r . As discussed before, the literature typically finds that reference points depend on expectations (e.g., Koszegi and Rabin, 2006; Abeler et al., 2011; Ahrens et al., 2014). Our experiment is designed such that \$0.15 constitutes the reference wage w^r (see section 2.1). As a result, in our experiment the labor supply curve derived above is kinked at the wage level of \$0.15.

⁶We only discuss the main implications of the model here since Ahrens et al. (2014) has all of the derivations.

⁷As before, we assume an upward sloping labor supply curve where the substitution effect dominates the income effect. That is, subjects work more when wages go up and they work less when wages fall. This assumption is also supported by our empirical findings.

4 Data and empirical approach

This section describes our outcome variables, details on the sample, and the empirical strategy used to identify the symmetry of wage effects.

4.1 Outcome Variables

We construct several outcome variables that measure different aspects of labor supply in order to identify the effect of wage changes on labor supply. These include the quit rate, number of transcribed pictures, time spent transcribing, transcription rate, and accuracy. Each of these variables is described in greater detail below.

Transcriptions and Hours Because workers are paid for each transcribed image, we expect that they will respond to the wage changes by changing the number of images they transcribe. Therefore, one variable of interest is the total number of transcribed images per worker. We further explore the the total time spent working on the task and the time per transcribed text (transcription rate). Because we do not have an exact measure of the time workers actually spent working on a picture, we proxy the transcription rate by counting the time between the submission of two transcriptions. We acknowledge that this likely overstates the transcription time for any given image. However, the difference in transcription rate between groups should still be instructive of the impact of wage changes.

Extensive Margin Recall that workers are notified of treatment after transcribing six images. The notification makes it clear that the worker has completed the HIT, but that there are additional (optional) images to transcribe. Workers are also informed that they can quit the task at this point or continue transcribing the additional images at the newly announced wage rate. Given these features of the treatment notification, we interpret the decision to stop working at this point as an extensive margin decision. Therefore, one of our key outcome variables is the share of workers who quit the task immediately after receiving the notification. Because the treatment notification has a modest nudge to quit, we expect that the share of quitters will be reasonably high in the control group despite the fact that the wage remains constant. The important question for us is: does the wage increase/decrease have any effect beyond this modest nudge.

An important feature of online-labor markets such as mTurk is that they facilitate almost instantaneous switching of labor tasks. In other words, a worker can quit one job this second and start a new job the next second. This is not unlike what one would observe in traditional labor markets where a worker secures a new job before quitting her existing job. Unfortunately, we do not observe what subjects do when they quit our task.

Therefore, the extensive margin response in our study simply means that the worker quits our task. We cannot say whether or not they quit working online or switch to a more profitable task. The most reasonable assumption, though, is that they simply switch to another task.

Accuracy Recall that the transcriptions are based on text for which we have the original digital copy. This makes it possible for us to measure accuracy by comparing the transcribed text for each worker to the actual text.

4.2 Sample

Our HIT was accepted by 1,168 mTurk workers. We restrict the sample to those workers who completed at least one picture, and therefore received the participation fee; this leaves us with 1,158 workers. We observe in the data that a few workers worked on the task for an unreasonable number of time, e.g., several days. To avoid this source of noise, we drop the top 0.05% of workers in the distribution of minutes worked; these are six workers who worked for more than 385 minutes on the task. Table 1 presents summary statistics for our sample of workers ($N = 1,152$) with regard to our main variables: number of transcribed pictures, accuracy of transcription, and total time worked. We observe that, on average, workers transcribed 12.8 pictures⁸ over an average time span of 39.79 minutes. The transcription quality was very high with an average accuracy of 96.97%. This is reassuring as it suggests that workers take the task seriously and provided high-quality transcriptions. Note that we intended to avoid giving the impression that subjects are participating in an experiment, and therefore did not survey any demographic characteristics.

Because the treatment variation in wages only appear after the first batch of six transcriptions, only a share of the total 1,152 participants are exposed to the treatment condition. Table 2 shows that 62.5% (720) of the 1,152 workers completed at least six pictures and therefore saw the treatment notification. This share ranges from 59% in the wage-increase group to 65% in the wage-decrease group. The number of observations in each treatment group is summarized in Table 2. In total, we have 248, 215, and 257 workers who saw the treatment notification in the control, increase and decrease groups, respectively. Because workers did not know they were in an experiment or that the wage rate would change, self-selection into the treatments was impossible. We therefore argue that the groups are balanced with respect to the characteristics that predict the probability of quitting before seeing the treatment, and thus we restrict the empirical analysis that follows to the sample of 720 participants who saw the treatment.

A common feature of mTurk is that workers discuss HITs on forums. This can

⁸Figure 14 in the Appendix provides the distribution of completed pictures for all workers in the sample.

raise issues for experimenters as those workers who have completed the experiment will unknowingly share the details of treatments with other workers who have yet to complete the experiment. We followed the forums on mTurk in order to determine if our HIT was being discussed and discovered that our HIT did in fact show up on one of the forums.⁹ The first mention of our HIT occurred on July 24 during the second run of the experiment. We noticed the mention on the 26th when the HIT had already expired. The discussion on the forum was favorable towards our HIT, but workers discussed the fact that the wage rate changed as well as the magnitude of the changes. They also discussed potential reasons for rate changes, and mostly speculated that the wage variation must be due the quality of work. Nobody speculated that this task is an experiment; people therefore still did not know they were part of an experiment.

The forum post led to a significant spike in acceptance of our HIT; approximately 58% of the workers accepted the HIT after the forum discussion began. Because some of these subjects knew of a potential wage variation before accepting the HIT, self-selection might be a problem. For example, it is possible that only workers who are willing to work for our lowest wage rate accepted our HIT. If this is the only source of selection, then our analysis produces a lower bound estimate in both groups. A more troubling source of selection is a case where workers sign up with the hope of receiving a wage increase. These subjects would effectively have a reference wage of \$0.18, and would be more likely to quit the task if assigned to the wage decrease group. This source of selection would lead to a downward bias in the wage-decrease group and upward bias in the wage-increase group. Because of this potential problem, we present estimates with and without the post-forum sample. There is no evidence that the forum had an effect on the results (see Appendix B).

4.3 Empirical Strategy

Random assignment to treatment groups ensures that our empirical approach is straight forward. We use non-parametric Wilcoxon rank-sum tests for differences in distributions between groups (Wilcoxon, 1945; Mann and Whitney, 1947). In addition, we run simple OLS regressions of the outcome variables on the treatment dummies. These empirical analyses allow us to identify the effect of wage changes on our outcome variables and to determine whether these responses are symmetric or not. To test for symmetry, we use the coefficients of the OLS regressions and t-tests to test the null that the sum of the estimated coefficients on the treatment dummies is zero.

The estimated treatment effects are then used to calculate implied elasticities separately for each treatment group. Using the control group as a counterfactual, we

⁹See https://www.reddit.com/r/HITsWorthTurkingFor/comments/3eg391/us_transcribe_texts_from_an_image_payment_bonus/.

derive the elasticity of an outcome variable Y with respect to wage for each treatment group i (either wage increase or decrease) as follows:

$$\epsilon_i = \frac{(Y_i - Y_c)/Y_c}{(w_i - w_c)/w_c}, \quad (1)$$

where subscript c indicates the control group, Y is the group average in the respective outcome variable, w is the wage per transcribed picture, $(w_i - w_c)$ is the change in wages in group i (either +3 or -3), and $(Y_i - Y_c)$ is the difference between the outcome variable in group i and the control group. Specifically, $(Y_i - Y_c)$ is the difference in means between the relevant treated group and the control group or, equivalently, the regression coefficient of the respective treatment dummy. Statistical significance of the elasticity is the same as the statistical significance of the respective treatment dummy in the regressions.

5 Results

We present the empirical results in this section. The mean of each outcome variable, by treatment group, is presented in Figures 6 to 13, and Table 3 shows the results of OLS regressions.

5.1 Extensive Margin

Figure 6 displays the treatment effects on the extensive margin, i.e., the share of workers who quit immediately after having seen the treatment variation. We observe that 14.1% of all workers in the control group quit the labor task after receiving the treatment notification. Relative to the control group, the share of quitters is 8.5 percentage points lower in the wage-increase group and 17.8 percentage points higher in the wage-decrease group. These group-wise differences between means are all statistically significant at the 1% level according to non-parametric ranksum tests. These results are also demonstrated in OLS regressions of the extensive-margin indicator variable on the treatment dummies; see Model I of Table 3.

An important observation is that the extensive margin response is asymmetric; the treatment effect for the wage-increase group is economically and statistically different from that for the wage-decrease group (p -value: 0.094). The asymmetry is also evident in the implied elasticities (as calculated by equation 1), which is 3.0 in the increase group and 6.3 in the decrease group.

5.2 Time responses

Time Spent Working. Figure 7 shows that, on average, subjects in the control group spent about 61 minutes working on the labor task. Relative to the control group, workers who experienced a wage increase worked on the task for 6 additional minutes while those who experienced a wage decrease spent 11 fewer minutes working on the task. A nonparametric test shows that the treatment effect is statistically different from zero for the wage-decrease group, but not for the wage increase group.

The nonparametric results are also reflected by the regressions; see Model III of Table 3. The differences indicate an asymmetric effect; the treatment effect is larger in the wage-decrease group than in the wage-increase group. This is also evident by the implied elasticities, which are 0.50 in the increase group and 0.87 in the decrease group. However, we cannot reject the null that the difference between the treatment effects is zero. In other words, though the relative magnitude of the treatment effects is indicative of an asymmetric response, we cannot rule out symmetry in a statistical sense.

The effect on the total time spent working described above can be decomposed into two parts; the first due to the extensive margin response and the second due to the intensive margin response. We identify the contribution of the intensive margin response in Figure 8, which plots the mean of the total number of minutes worked conditional on *not* quitting right away after the treatment. The Figure shows that, conditional on transcribing at least one picture after the treatment notification, workers in the control group spent an average of 68 minutes on the task. Relative to the control group, workers in the wage-increase group worked for one additional minute while workers in the wage-decrease group spent 4 fewer minutes on the task. Subtracting these intensive-time-margin treatment effects from the total treatment effects implies that the extensive margin response explains the overwhelming majority of the effect on time spent working on the task. In fact, the extensive margin response explains 83% ($= (6 - 1)/6$) and 64% ($= (11 - 4)/11$) of the time margin response in the wage increase and decrease groups, respectively.

Transcription Rate. The results for the transcription rate are shown in Figure 9. Workers on average spent 3.8, 3.4 and 3.9 minutes for one picture in the control, increase and decrease groups, respectively. The differences between groups are not statistically significant (also see Model IV in Table 3). We further separate this total effect into its intensive and extensive margin components and find that there is no statistically significant effect on either margin (see Figure 10 which reports the transcription rate conditional on completing at least one transcription after the treatment notification).

5.3 Number and quality of transcriptions

Number of Transcribed Pictures. Figure 11 shows that the treatment variation clearly affected the number of transcribed pictures per worker. While the average worker transcribed 19.04 images in the control group, the average worker completed 22.35 and 15.25 pictures in the wage-increase and wage-decrease groups, respectively. All group-wise differences between groups are distinguishable from zero at the 1%-level according to non-parametric rank-sum tests. These results are confirmed in Model II of Regression Table 3, which also shows that we cannot reject the null that the wage effect on total output is symmetric.

As in section 5.2, we decompose the total effect on number of transcribed pictures into its intensive and extensive margin components. We begin with the contribution of the intensive margin response by calculating the per-worker number of transcriptions for each group conditional on completing at least one picture after seeing the treatment information. These results, which are presented in Figure 12, show that output is higher when wages rise and lower when wages fall. While the non-parametric tests reveal that the difference between the control and increase group is statistically significant, the difference between control and decrease is not significant (p-value: 0.15). More importantly, the magnitude of these intensive-time-margin effects is not asymmetric in a statistical sense.

We next identify the contribution of the extensive margin response by subtracting the intensive-time-margin effect from the total effect. For example, the total treatment effect for the wage-increase group is 3.3 transcriptions. From Figure 12, we know that 2.14 of this effect is due to the intensive-time-margin response. Therefore, the balance of 1.17 ($= 3.31 - 2.14$) is due to the extensive margin response. A similar calculation for the wage-decrease group reveals that the contribution of the extensive margin is 2.19. The fact that 2.19 is almost twice as large as 1.17 suggests that the contribution of the extensive margin response is asymmetric.

Quality of transcriptions. Figure 13 depicts that the wage-rate changes did not have any effects on the quality of transcription. The differences are tiny and indistinguishable from zero, which confirms that workers in all groups worked paid careful attention to the task. This result is in line with the field experiment of Kube et al. (2013) who do not find any effects of wages on work quality either.

5.4 Robustness

Because the workers discussed our task on the mTurk forum, it is possible that our findings are driven by selection into the HIT. We explore this by performing the analyses separately on the sample of workers who worked on our task before it was discussed online and the

sample of workers who worked on it afterwards. These results, which are presented in Appendix B, show no evidence that our results are driven by selection among workers who participated in the post-forum period. In addition, we regress each outcome variable on a dummy variable indicating whether the subject worked on the task before or after the forum post; we do not find any significant effects of working on the task after the treatment (results not reported).

6 Discussion of results

We begin this section by arguing that our findings are due to reference-dependent utility functions with loss aversion. This is followed by a discussion of other possible explanations for our findings; in each case we argue that the alternatives are less likely than loss-aversion. We then describe the policy implications and generalizability of our findings.

6.1 Mechanisms

Loss Aversion. Our results show that the extensive margin response to wage changes is strongly asymmetric. We also find evidence of an asymmetric intensive-time-margin response, but this effect is not statistically distinguishable from zero. Finally, the wage-induced effect on number of transcribed images is symmetric. Interestingly, the contribution of the extensive margin response to the observed changes in transcription is asymmetric. Over all, our results point to an upward sloping supply curve that appears to be kinked around the reference wage of \$0.15.

We argue that our findings can be rationalized by the theoretical model put forward by Ahrens et al. (2014); see section 3. Workers are loss averse with reference-dependent utility functions. This implies that workers' labor-supply functions are kinked at \$0.15, and have steeper slopes in the gain domain than in the loss domain. In this framework, a \$0.03 wage decrease is predicted to have a larger labor-supply effect than a \$0.03 wage increase. Our findings are consistent with this prediction. Our findings are also consistent with the empirical results of Kube et al. (2013). Importantly, our results add to these two papers by further illuminating the nature of the asymmetry of labor supply responses to wage changes. In particular, we find that asymmetry is more pronounced on the extensive margin relative to the intensive margin. This refinement of the asymmetry is especially important since it is generally accepted that labor supply elasticities are mainly determined by the extensive margin response (Blundell and MaCurdy, 1999; Meghir and Phillips, 2010; Bargain et al., 2014). That we find evidence for an upward sloping labor supply curve is also consistent with the labor supply literature (e.g., Keane, 2011; Bargain et al., 2014).

Standard Labor Supply. To argue that our asymmetric results are due to loss aversion requires that we rule out rational responses predicted by the standard model. For example, one possible explanation of our extensive-margin results is that they are driven by a rational response to the difference between reservation wages and the newly announced wage. The argument goes as follows. A worker’s decision to work or not is determined by the wage rate relative to the worker’s reservation wage. The worker chooses to work if the wage rate is greater than her reservation wage. Since participation in our experiment is voluntary, it is reasonable to assume that the reservation wage for our workers has a distribution that is bounded between \$0 and \$0.15. This raises the possibility that some workers have reservation wage between \$0.12 and \$0.15. If true, this would make the observed responses consistent with a rational calculus instead of behavioral biases. In particular, we would expect all rational workers with reservation wage between \$0.12 and \$0.15 to quit the labor task when the wage rate decreases to \$0.12. We argue that there are at least two reasons to rule out this possible explanation. First, using a labor task that is very similar to ours, Horton et al. (2011) find that mTurkers in their experiment had an implicit median reservation wage of only \$0.14 per hour, which is substantially lower than the implied hourly wage of \$1.90¹⁰ in our wage decrease group.¹¹ Second, we observe a statistically significant extensive margin response in the wage-increase group, which cannot be explained by reservation wage argument since every worker in this group would have been paid above the reservation wage from the beginning of the experiment. This, combined with the fact that the responsiveness of the wage-decrease group is approximately twice that of the wage-increase group, suggests that the difference between reservation wage and announced wage rate is a very unlikely explanation of our extensive-margin results.

So what about the intensive margin results? Could these results be due to the standard model. We argue that this is also unlikely. Notice that the intensive margin response is based on the difference between the marginal disutility of transcription and the wage rate. Assuming the disutility of transcription is increasing in the number of transcriptions, we would expect workers in the wage-increase group to work longer and faster, relative to the control group. On the other hand, because the wage-decrease group faces a lower fixed wage than the control group, we would expect workers in this group to spend less time working and to do so at a slower rate. This is exactly what we find. However, contrary to the symmetric response predicted by the standard model, we find that the economic magnitude of these responses is asymmetric; e.g., the intensive-margin

¹⁰Note that this hourly wage of \$1.90 is a lower bound because our measure of the time it takes to transcribe one picture overstates the actual time per picture; see section 4.1.

¹¹Horton et al. (2011) estimated the reservation wage using data generated from an mTurk task. The task required mTurk workers to transcribe paragraph-sized chunks of text that is written in Tagalog, a language of the Philippines. That is, as in our task, subjects are required to transcribe foreign language text and are paid per transcribed text.

treatment effect for the time spent working is in the wage-decrease group is four times that in the wage-increase group.

We acknowledge that standard neoclassical labor supply model may yield asymmetric labor supply responses, but only if the labor supply function has a particular shape. Even then, asymmetry would only arise for non-marginal wage changes. Although our wage rate changed by 20% from the reference point, the absolute change is only \$0.03, which is small, and may be viewed as a rather marginal change. We argue then, that we have a “very small” wage change, which rules out the standard model as a possible explanation for our results.

Reciprocity. Another possible explanation of our findings is reciprocity; workers interpret the wage changes as punishment or reward, and respond accordingly. Workers who receive a wage decrease feel punished and thus lower their labor supply in an effort to punish the employer, while workers who receive a wage increase feel rewarded for their effort and thus work harder to return the favor to their employer. To the extent that the degree of induced reciprocity is asymmetric around a reference wage, this explanation is potentially consistent with reference-dependence and loss-aversion, and therefore is a potential driver of our findings. Although we have no way of ruling out this motivation behind our results, we argue that this is an unlikely explanation based on our experimental design. Recall that subjects are paid for each completed transcription and not per-unit-of-time. This implies that workers in our experiment have little scope for punishing the employer through shirking. Additionally, reducing the number of transcription implies that a worker punishes herself in the form of lower pay-off, and potentially lower performance rating, which affects her prospects of being allowed to work on other mTurk tasks.¹² One strategy to punish the employer without incurring a cost is to continue to work hard, but submit transcriptions that are of low enough quality to be mostly useless to the employer, but high enough quality to avoid a negative performance review. Because we have the transcriptions and the actual texts, we can check to see if workers used this strategy; there is no evidence that they did (see section 5.3).

Similarly, as opposed to settings where workers are paid per hour, transcribing more pictures is not a reward for the employer in our experiment since this increases the costs to the employer. Workers are also likely to know that employers can easily recruit other workers to transcribe pictures and that employers therefore do not face the risk that pictures remain untranscribed.

¹²Workers on mTurk receive performance rating for each task they complete. Employers often use workers’ performance rating to screen out low performers from their tasks. Therefore, a worker who decides to punish us because their wage has been reduced, runs the risk of limiting the number of tasks she will qualify to work on in the future.

6.2 Implications

The existing labor supply literature identifies labor supply elasticities by exploiting data comprised of both wage increases and decreases. This approach makes sense in the context of the standard labor supply model where the elasticity is shown to be symmetric. It also makes sense when one considers that nominal wages are almost always rising. However, this approach becomes problematic if one believes that workers respond to real wages rather than nominal wages, and that workers have reference-dependent utility functions. The reason is that real wages vary more greatly over time and generally reflect both increases and decreases. As Ahrens et al. (2014) have shown theoretically and we have found empirically, labor supply responds asymmetrically to wage changes under these circumstances. Our results suggest that this is especially true for the extensive margin responses that drive labor supply elasticities. Relying on the standard estimation approach under these circumstances leads to overestimated elasticities when wages rise and underestimated elasticities when wages fall.

From a purely academic perspective, our findings confirm that labor supply is best modeled with reference-dependent utility functions that allows the modeler to account for an asymmetric response to wages. Our findings are also practically useful, since labor supply elasticities play an important role in quantifying the economic impacts of policy changes that affect wages; e.g., minimum wage policies. One policy area where our findings are likely to be particularly useful is taxation. Tax reforms generally result in either tax increases or tax decreases. In fact, upward and downward changes are more prominent for tax rates than for wages. Further, the tax elasticity of labor supply plays a key role in determining the efficiency cost and revenue impacts of tax policy changes. We know that the tax elasticity of labor supply is generally larger than the wage elasticity: e.g., due to tax aversion (Kessler and Norton, 2015). This suggests that the labor-supply asymmetry with respect to tax-rate changes is likely to be more pronounced than what our findings for labor supply responses to wage changes suggest. This makes the distinction between rate increases and decreases more particularly important in the context of tax rate.

6.3 Generalizability

The results described above are obtained using an experimental design in a large real-world labor market. Importantly, workers did not know they participated in an experiment and thus behaved as they would in their natural occurring environment. Due to randomization, our experimental design also guarantees internal validity. Though our findings are based on an actual real-world labor market, we are careful not to generalize our results to all types of labor markets. Nonetheless, we argue that the findings are applicable to any labor market with piece rate, flexibility and multiple outside options. One example of such

labor markets is on-line crowd-sourcing labor markets, which are becoming increasingly common in the current technological age.¹³ A common feature of these labor markets is that workers tend to work for relatively low wages and have extremely high levels of labor supply flexibility. Because the labor supply effects are predominately on the extensive margin, we argue that the results are also likely to be equally applicable to traditional labor markets where workers face greater restrictions on labor hours.

7 Conclusion

We estimate the effect of wage change on labor supply using data generated in a field experiment. We find strong evidence of an asymmetric response on the extensive margin. The magnitude of the intensive-time margin responses is also indicative of an asymmetric response, but we cannot rule out symmetry in a statistical sense. Though we cannot rule out a symmetric response in the number of transcribe images, the evidence does suggest that the contribution of the extensive margin to the effect of wages on transcriptions is asymmetric. Overall, our findings suggest that the labor supply curve for mTurkers is upward sloping, and is best modeled by a reference-dependent utility function that accommodates loss aversion. In our particular setting, we find that the supply curve is upward sloping with a kink at a wage rate of \$0.15. We speculate that a similar, but much larger asymmetric response exist for the tax elasticity of labor supply.

¹³See <https://sites.google.com/site/johnjosephorton/miscellany/online-labor-markets> for a list of crowd-sourcing online labor markets.

References

- Abeler, J., Falk, A., Goette, L. and Huffman, D. (2011). Reference points and effort provision, *American Economic Review* **101**(2): 470–92.
- Ahrens, S., Pirschel, I. and Snower, D. J. (2014). A theory of wage adjustment under loss aversion, *IZA Discussion Papers 8699*, Institute for the Study of Labor.
- Bargain, O., Orsini, K. and Peichl, A. (2014). Comparing Labor Supply Elasticities in Europe and the United States – New Results, *The Journal of Human Resources* **49**(3): 723–838.
- Blundell, R. and MaCurdy, T. E. (1999). Labor Supply: A Review of Alternative Approaches, in O. Ashenfelter and D. Card (eds), *Handbook of Labor Economics*, North-Holland, Amsterdam.
- Buhrmester, M., Kwang, T. and Gosling, S. D. (2011). Amazon’s mechanical turk: A new source of inexpensive, yet high-quality, data?, *Perspectives on Psychological Science* **6**(1): 3–5.
- Camerer, C., Babcock, L., Loewenstein, G. and Thaler, R. (1997). Labor supply of new york city cabdrivers: One day at a time, *The Quarterly Journal of Economics* **112**(2): 407–441.
- Card, D. and Dahl, G. B. (2011). Family violence and football: The effect of unexpected emotional cues on violent behavior, *The Quarterly Journal of Economics* **126**(1): 103–143.
- Crawford, V. P. and Meng, J. (2011). New york city cab drivers’ labor supply revisited: Reference-dependent preferences with rational-expectations targets for hours and income, *American Economic Review* **101**(5): 1912–32.
- Dunn, L. F. (1996). Loss aversion and adaptation in the labor market: Empirical indifference functions and labor supply, *The Review of Economics and Statistics* **78**(3): 441–450.
- Eissa, N. and Liebman, J. B. (1996). Labor supply response to the earned income tax credit, *The Quarterly Journal of Economics* **111**(2): 605–637.
- Evers, M., De Mooij, R. and Van Vuuren, D. (2008). The wage elasticity of labour supply: A synthesis of empirical estimates, *De Economist* **156**(1): 25–43.
- Fehr, E. and Goette, L. (2007). Do workers work more if wages are high? evidence from a randomized field experiment, *American Economic Review* **97**(1): 298–317.

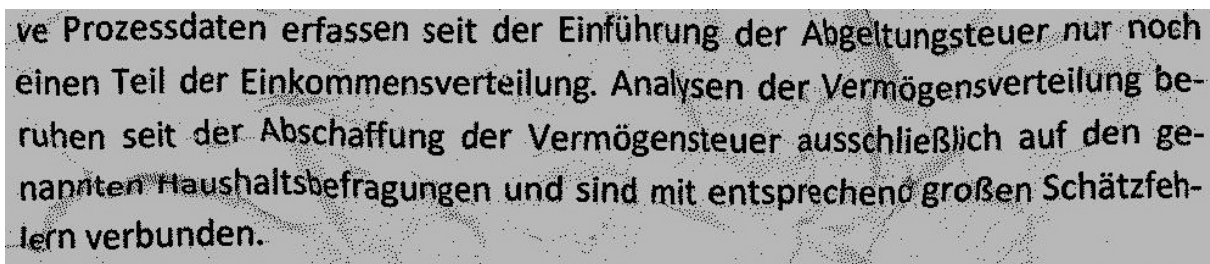
- Goette, L., Huffman, D. and Fehr, E. (2004). Loss aversion and labor supply, *Journal of the European Economic Association* **2**(2-3): 216–228.
- Horton, J. J., Rand, D. G. and Zeckhauser, R. J. (2011). The online laboratory: conducting experiments in a real labor market, *Experimental Economics* **14**: 399–425.
- Kahn, S. (1997). Evidence of nominal wage stickiness from microdata, *The American Economic Review* **87**(5): 993–1008.
- Kahneman, D. and Tversky, A. (1979). Prospect Theory: An Analysis of Decision under Risk, *Econometrica* **47**(2): 263–91.
- Keane, M. P. (2011). Labor supply and taxes: A survey, *Journal of Economic Literature* **49**(4): 961–1075.
- Kessler, J. B. and Norton, M. I. (2015). Tax aversion in labor supply, *Journal of Economic Behavior and Organization* **forthcoming**.
- Kleven, H. J. and Schultz, E. A. (2014). Estimating taxable income responses using danish tax reforms, *American Economic Journal: Economic Policy* **6**(4): 271–301.
- Koszegi, B. and Rabin, M. (2006). A model of reference-dependent preferences, *The Quarterly Journal of Economics* **121**(4): 1133–1165.
- Koszegi, B. and Rabin, M. (2007). Reference-dependent risk attitudes, *American Economic Review* **97**(4): 1047–1073.
- Koszegi, B. and Rabin, M. (2009). Reference-dependent consumption plans, *American Economic Review* **99**(3): 909–36.
- Kube, S., MarÅ©chal, M. A. and Puppe, C. (2013). Do wage cuts damage work morale? evidence from a natural field experiment, *Journal of the European Economic Association* **11**(4): 853–870.
- Kuziemko, I., Norton, M. I., Saez, E. and Stantcheva, S. (2015). How elastic are preferences for redistribution? evidence from randomized survey experiments, *American Economic Review* **105**(4): 1478–1508.
- Mann, H. B. and Whitney, D. R. (1947). On a test whether one of two random variables is stochastically larger than the other, *Annals of Mathematical Statistics* **18**: 50–60.
- Marzilli Ericson, K. M. and Fuster, A. (2011). Expectations as endowments: Evidence on reference-dependent preferences from exchange and valuation experiments, *The Quarterly Journal of Economics* **126**(4): 1879–1907.

- Mason, W. and Suri, S. (2011). Conducting behavioral research on amazon's mechanical turk, *Behavioural Research* **44**: 1–23.
- Meghir, C. and Phillips, D. (2010). Labour Supply and Taxes, *in* J. Mirrless, S. Adam, T. Besley, R. Blundell, S. Bond, R. Chote, M. Gammie, P. Johnson, G. Myles and J. Poterba (eds), *Dimensions of Tax Design: The Mirrlees Review*, Oxford University Press, pp. 202–274.
- Paolacci, G., Chandler, J. and Ipeirotis, P. G. (2010). Running experiments on amazon mechanical turk, *Judgment and Decision Making* **5**(5).
- Pope, D. G. and Schweitzer, M. E. (2011). Is tiger woods loss averse? persistent bias in the face of experience, competition, and high stakes, *American Economic Review* **101**(1): 129–57.
- Post, T., van den Assem, M. J., Baltussen, G. and Thaler, R. H. (2008). Deal or no deal? decision making under risk in a large-payoff game show, *American Economic Review* **98**(1): 38–71.
- Rothstein, J. (2010). Is the eite as good as an nit? conditional cash transfers and tax incidence, *American Economic Journal: Economic Policy* **2**(1): 177–208.
- Saez, E., Slemrod, J. and Giertz, S. H. (2012). The elasticity of taxable income with respect to marginal tax rates: A critical review, *Journal of Economic Literature* **50**(1): 3–50.
- Wilcoxon, F. (1945). Individual comparisons by ranking methods, *Biometrics* **1**: 80–83.

8 Tables and Figures

8.1 Figures

Figure 1: Image of Text to be Transcribed



ve Prozessdaten erfassen seit der Einführung der Abgeltungsteuer nur noch einen Teil der Einkommensverteilung. Analysen der Vermögensverteilung beruhen seit der Abschaffung der Vermögensteuer ausschließlich auf den genannten Haushaltsbefragungen und sind mit entsprechend großen Schätzfehlern verbunden.

Notes: The Figure depicts a screenshot of an image of text that was to be transcribed by the subjects. Subjects were randomly shown one of 128 images. All images are comparable to the image depicted in the Figure. All images are in German and taken from a recent policy-report publication.

Figure 2: Human Intelligence Task Shown on mTurk

Transcribe texts from an image. Payment: bonus of \$0.15 per transcribed image plus a flat reward for completing the HIT
Requester: ML Reward: \$0.10 per HIT HITs Available: 1 Duration: 4 hours
Qualifications Required: Location is US

Instructions

This hit requires you to transcribe texts which have been scanned from an old German document (see external link for an example). You can transcribe as many of the texts as you want. You will be paid 0.15 USD bonus for each transcribed text. In addition, you receive the 0.10 USD reward for working on this HIT (this reward is paid once and not for each text). You only get paid if you transcribe at least one picture. Transcriptions will be checked for accuracy before bonus is paid.

Please read the instructions carefully.

Instructions

1. Go to the external website shown below and transcribe as many pictures as you want.
2. You will be shown a text and an empty text box. Please complete your transcription of the text in the text box.
3. Please use the following rules for non-standard characters
 1. transcribe ä as ae, Ä as Ae
 2. transcribe ö as oe, Ö as Oe
 3. transcribe ü as ue, Ü as Ue
 4. transcribe ß as ss
4. If you cannot read some characters or you are unsure about them, please replace them with an underscore _.
5. Please press 'Save picture' after you are finished transcribing the text show on the page; the next text will be shown after you press 'Save picture'.
6. You can stop at any time. Please do not forget to copy your Personal ID to the textfield shown below before submitting and closing this HIT. Your Personal ID number is displayed in the top right corner of each page on the external website. Please note this Personal ID is not the same as your Amazon Worker ID.

Want to work on this HIT?

Notes: The Figure depicts a screenshot from the Amazon Turk website. It shows how the labor task used for the field experiment was advertised on Amazon Turk. Subjects are taken to our external website once they click the "Accept Hit" button.

Figure 3: Instructions Shown on Our Website

Transcribe pictures

Personal ID: 789db7d48af873208f7f253a6cd5a24c
Transcribed pictures: 0
Current bonus per picture: 0.15 USD

Welcome.

Thank you for working on this hit. This hit requires you to transcribe texts which have been scanned from an old German document (see below for an example). You can transcribe as many of the texts as you want; a new text will be presented when you hit the 'Save picture' button. You will be paid 0.15 USD bonus for each transcribed text. In addition, you receive the 0.10 USD reward as shown on the Amazon Mechanical Turk page for working on this HIT (this reward is paid once and not for each text). You only get paid if you transcribe at least one picture. Transcriptions will be checked for accuracy before bonus is paid.

To the top right of this web page you see your personal ID. Please submit this personal identifier to Amazon Mechanical Turk in order to complete this assignment.

Instructions

1. Your Personal ID number is shown in the top right corner of each page. Please submit this personal identifier to Amazon Mechanical Turk in order to complete this assignment.
2. You will be shown a text and an empty text box. Please complete your transcription of the text in the text box.
3. Please use the following rules for non-standard characters
 1. transcribe ä as ae, Ä as Ae
 2. transcribe ö as oe, Ö as Oe
 3. transcribe ü as ue, Ü as Ue
 4. transcribe ß as ss
4. If you cannot read some characters or you are unsure about them, please replace them with an underscore _.
5. Please press 'Save picture' after you are finished transcribing the text show on the page; the next text will be shown after you press 'Save picture'.
6. You can stop at any time. Please do not forget to copy your Personal ID to the Amazon Turk Website before submitting and closing this HIT

Notes: The Figure depicts a screenshot of the external website that we set up for the purpose of the field experiment. Subjects were taken to this website once they decided on the Amazon Turk website that they would like to work on the task. The depicted screenshots provides subjects all information relevant for the task.

Figure 4: Treatment Variation

Transcribe pictures

Personal ID: 789db7d48af873208f7f253a6cd5a24c
Transcribed pictures: 6
Current bonus per picture: 0.12 USD

Thank you for transcribing these pictures. As written in the introduction, we will grant a bonus of 0.15 USD for each of these pictures.

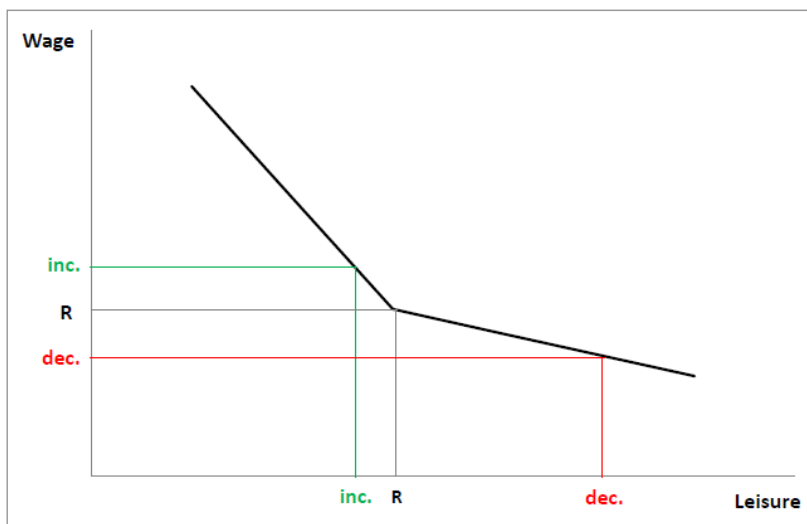
There are additional pictures that you can transcribe. However, the bonus payment for each additional picture will be 0.12 USD from now on. You will receive 0.15 USD bonus for each of the 6 pictures you transcribed so far, though.

If you want to stop and exit, just copy your Personal ID to the Amazon Turk Website and submit the HIT.

Continue

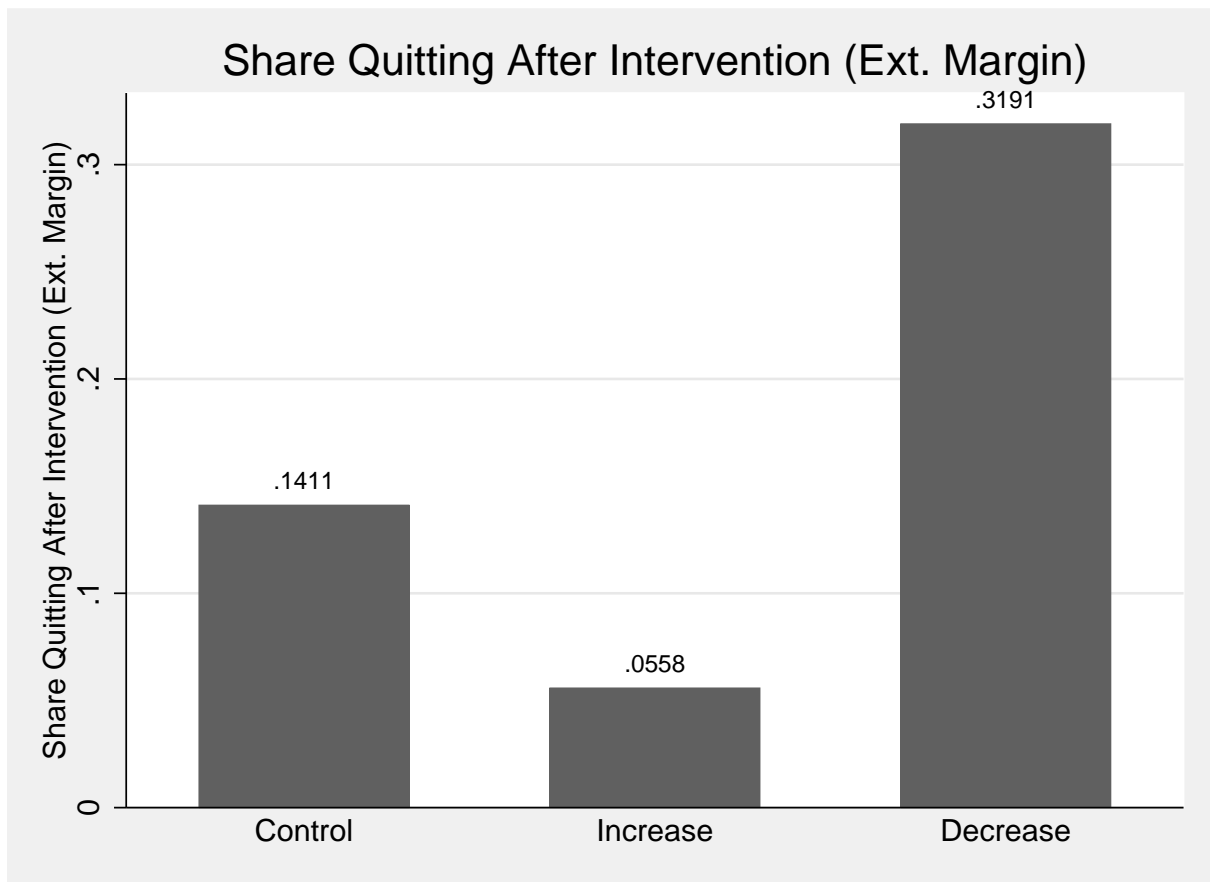
Notes: The Figure depicts a screenshot of the treatment notification in the "wage decrease" group. The treatment notifications for the "control" and "wage increase" groups were identical except for the information regarding the piece-rate wage for the subsequent images. The treatment notification popped up after a subject transcribed six images.

Figure 5: Prediction: Labor Supply under Loss Aversion



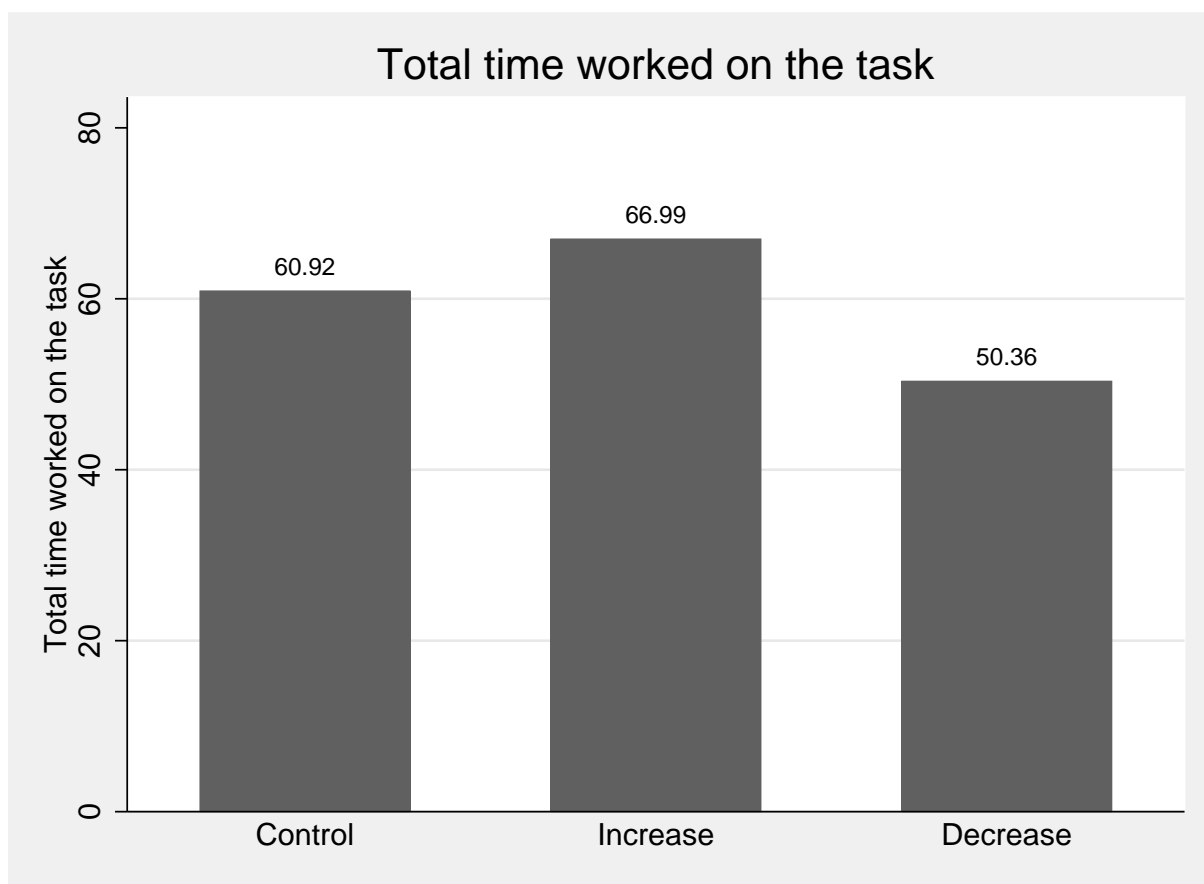
Notes: The Figure displays the relationship between leisure and wages under loss aversion. The curve is kinked at the reference wage R . Individuals who currently face the reference wage respond stronger to a wage decrease (by supplying less labor/more leisure) than to a wage increase of equal magnitude (by supplying more labor/less leisure).

Figure 6: Extensive Margin by Treatment Group



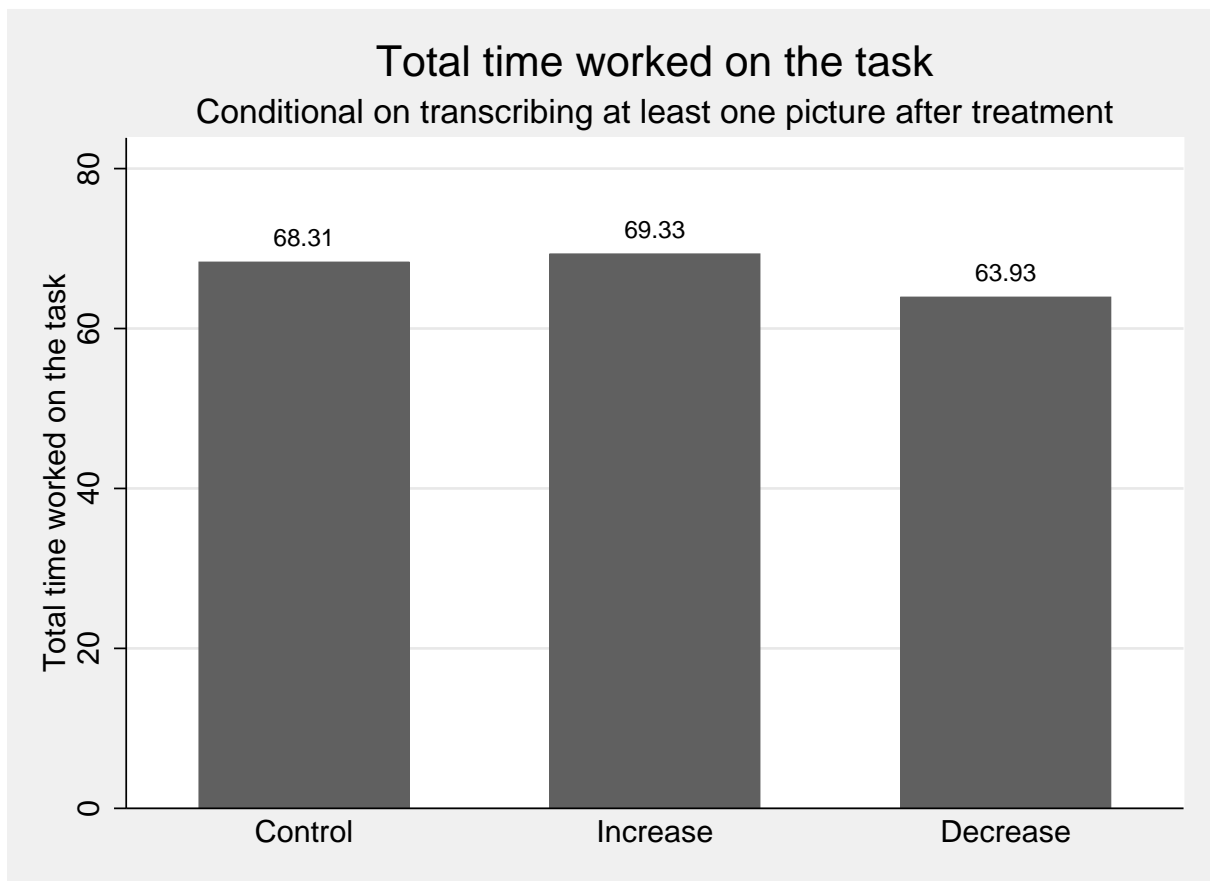
Notes: The Figure depicts the share of subjects in each group who quit the labor task immediately after seeing the treatment notification (i.e., share of subjects who transcribed six pictures but not a seventh one). The number of observations is 720 with 248 subjects in the control group, 215 subjects in the "wage increase" group and 257 subjects in the "wage decrease" group. All 720 subjects in the sample have transcribed at least six images.

Figure 7: Total Time Worked by Treatment Group



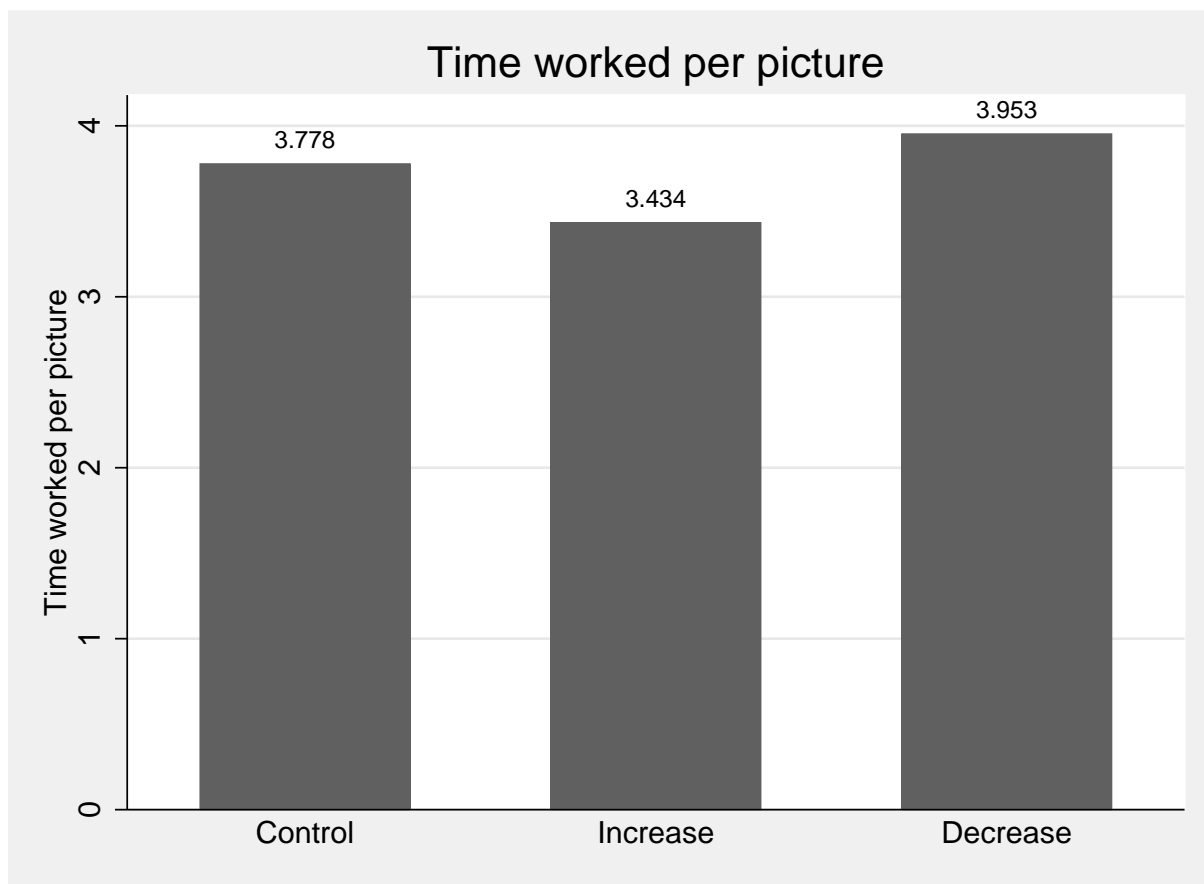
Notes: The Figure depicts for each group the average time (in minutes) that subjects totally spent on working on the labor task. The number of observations is 720 with 248 subjects in the control group, 215 subjects in the "wage increase" group and 257 subjects in the "wage decrease" group. All 720 subjects in the sample have transcribed at least six images.

Figure 8: Total Time Worked by Treatment Group: Intensive margin



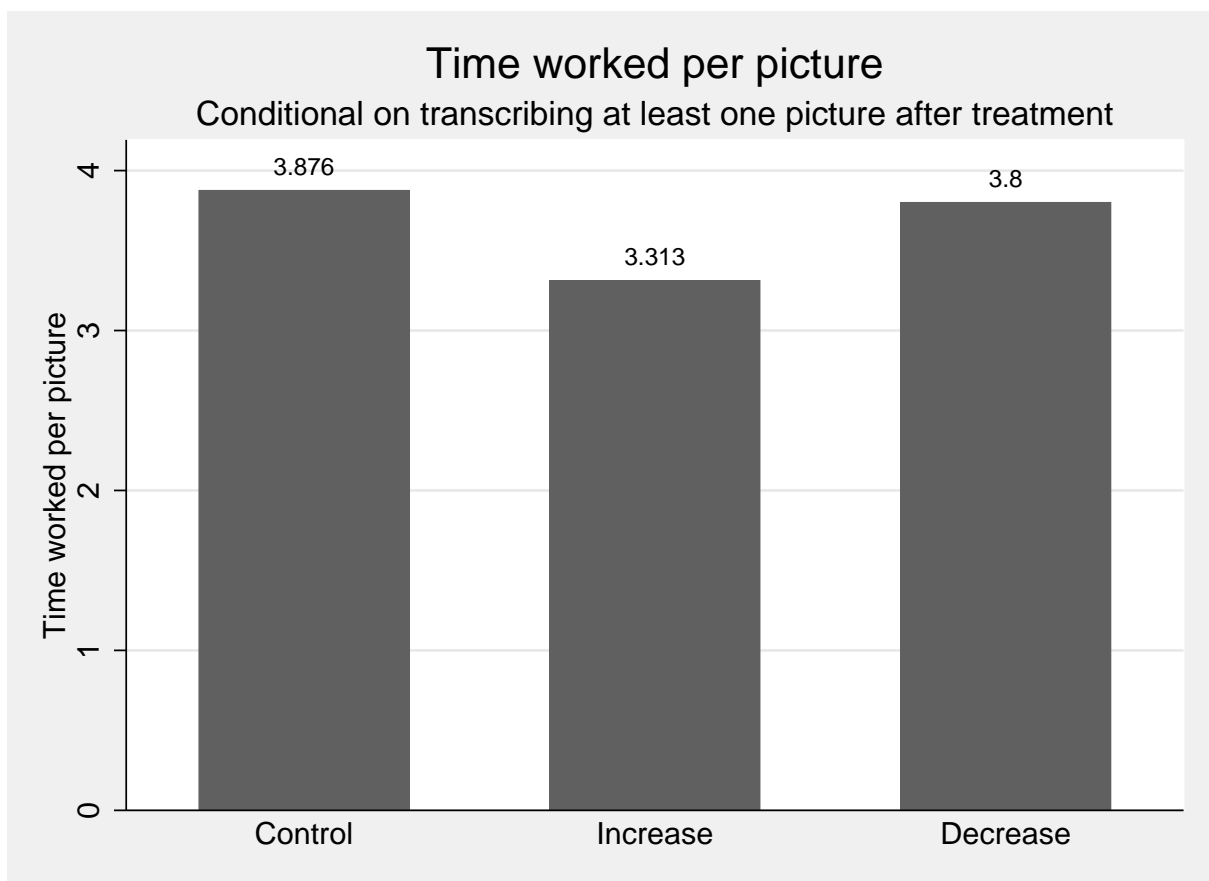
Notes: The Figure depicts for each group the average time (in minutes) that subjects totally spent on working on the labor task. The underlying sample is restricted to subjects who did *not* quit the labor task immediately after seeing the treatment notification (i.e., restricted to subjects who have transcribed at least seven images). The number of observations is 591 with 213 subjects in the control group, 203 subjects in the "wage increase" group and 175 subjects in the "wage decrease" group. All 591 subjects in the sample have transcribed at least seven images.

Figure 9: Avg. Time per Transcription by Treatment Group



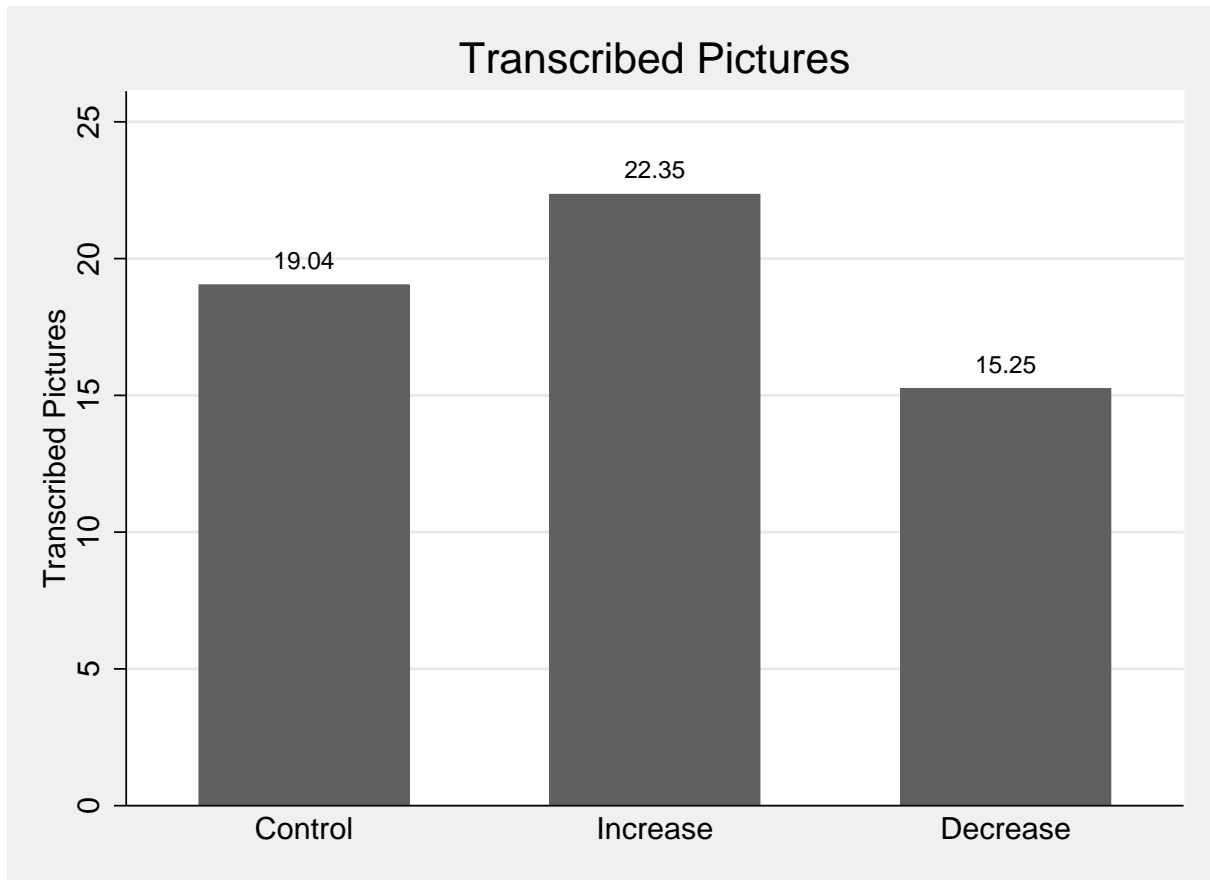
Notes: The Figure depicts for each group the average time (in minutes) that subjects spent working on one image. The number of observations is 720 with 248 subjects in the control group, 215 subjects in the "wage increase" group and 257 subjects in the "wage decrease" group. All 720 subjects in the sample have transcribed at least six images.

Figure 10: Avg. Time per Transcription by Treatment Group: Intensive margin



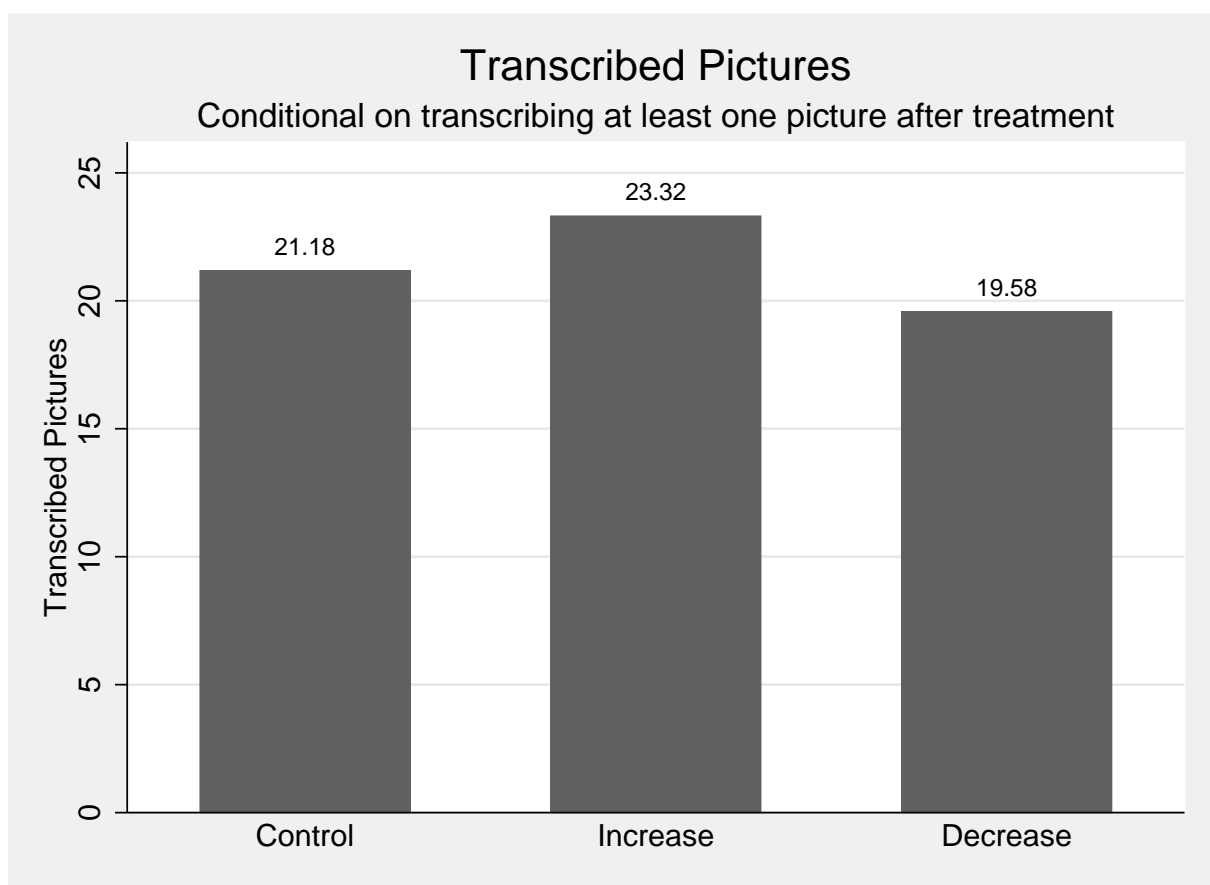
Notes: The Figure depicts for each group the average time (in minutes) that subjects spent working on one image. The underlying sample is restricted to subjects who did *not* quit the labor task immediately after seeing the treatment notification (i.e., restricted to subjects who have transcribed at least seven images). The number of observations is 591 with 213 subjects in the control group, 203 subjects in the "wage increase" group and 175 subjects in the "wage decrease" group. All 591 subjects in the sample have transcribed at least seven images.

Figure 11: Number of Transcribed Pics by Treatment Group



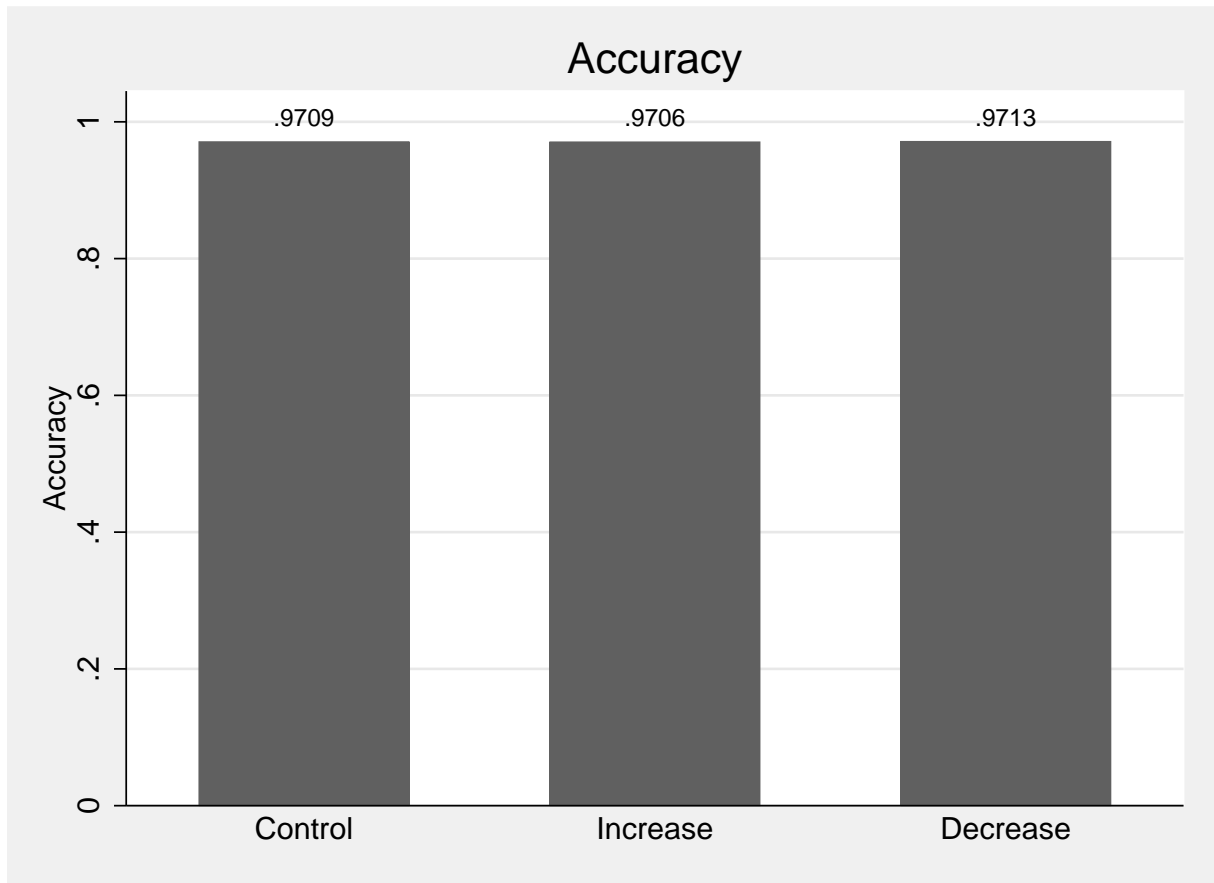
Notes: The Figure depicts for each group the average number of images that subjects transcribed. The number of observations is 720 with 248 subjects in the control group, 215 subjects in the "wage increase" group and 257 subjects in the "wage decrease" group. All 720 subjects in the sample have transcribed at least six images.

Figure 12: Number of transcribed pics conditional on workers who completed at least one pic after the treatment notification



Notes: The Figure depicts for each group the average number of images that subjects transcribed. The underlying sample is restricted to subjects who did *not* quit the labor task immediately after seeing the treatment notification (i.e., restricted to subjects who have transcribed at least seven images). The number of observations is 591 with 213 subjects in the control group, 203 subjects in the "wage increase" group and 175 subjects in the "wage decrease" group. All 591 subjects in the sample have transcribed at least seven images.

Figure 13: Accuracy by Treatment Group



Notes: The Figure depicts for each group the average transcription accuracy, i.e., the average share of characters in each image that is transcribed correctly. The number of observations is 720 with 248 subjects in the control group, 215 subjects in the "wage increase" group and 257 subjects in the "wage decrease" group. All 720 subjects in the sample have transcribed at least six images.

8.2 Tables

Table 1: Summary statistics: Pictures transcribed and Accuracy

variable	N	mean	sd	p10	p50	p90
Pics transcribed	1152	12.81	13.23	2.00	7.00	33.00
Total time	1152	39.79	50.25	3.17	19.64	104.35
Accuracy	1151	0.97	0.02	0.96	.97	0.98

Notes: Summary statistics for outcome variables. The sample is all subjects who started working on the task (i.e., including those who did not necessarily get to see the treatment notification after 6 transcribed pictures). *Pics transcribed* is the average number of images that subjects transcribed. *Total time* is the average time (in minutes) that subjects totally spent on working on the labor task. *Accuracy* the average share of characters that is transcribed correctly. *N* is the number of observations. *sd* is the standard deviation. *pX* indicates the X-th percentile.

Table 2: Number of Observations

Group	Seen Treatment		Total
	No	Yes	
Control	143	248	391
Increase	149	215	364
Decrease	140	257	397
Total	432	720	1152

Notes: Number of observations by treatment group who (i) started working on the task but did not see the treatment notification, i.e., they transcribed five images or less (Column *No*) and (ii) who transcribed at least six pictures and therefore saw the treatment notification (Column *Yes*).

Table 3: Regression estimates and implied elasticities

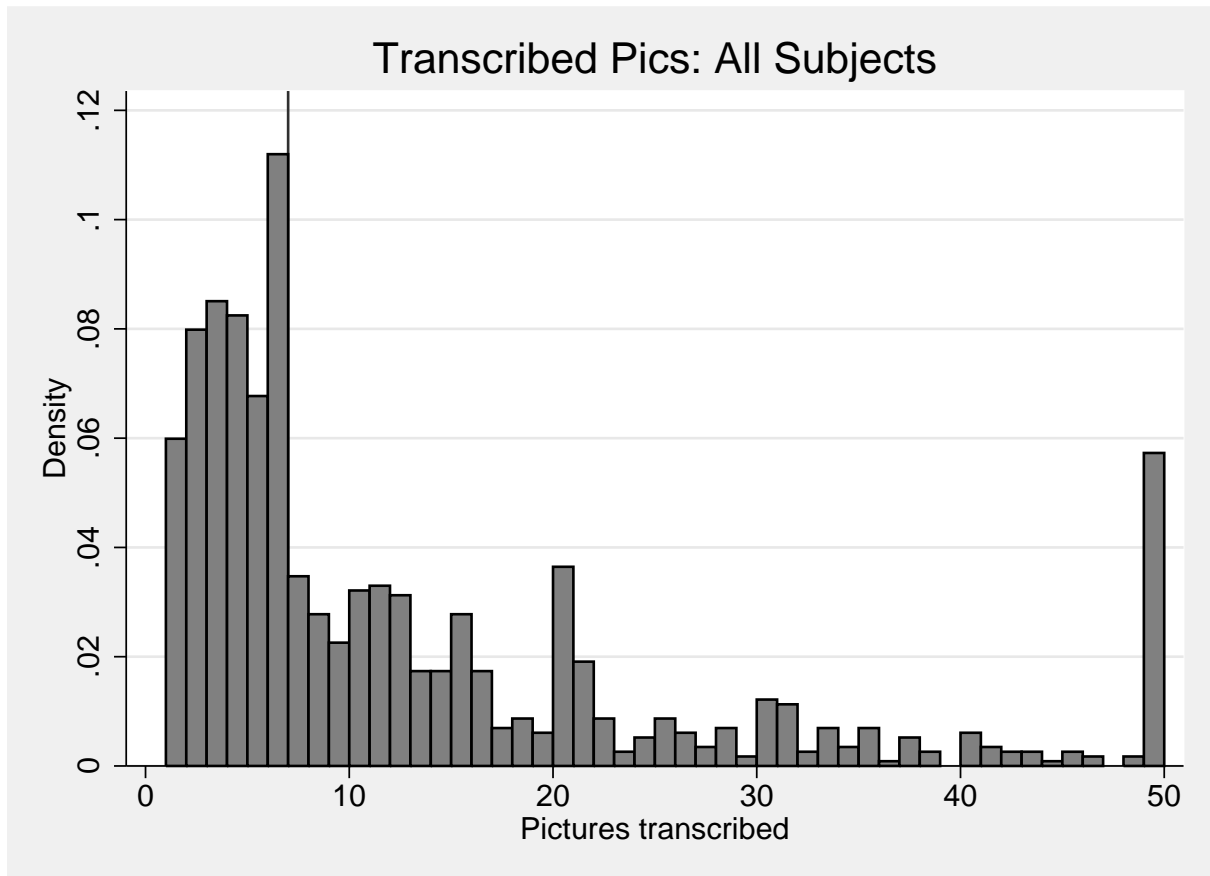
	Dependent Variable				
	I	II	III	IV	V
	Ext. margin	Pics	Total Time	Mean Time	Accuracy
<i>Reference group: Control</i>					
Increase	-0.085*** (0.027)	3.309** (1.298)	6.078 (4.985)	-0.344 (0.264)	-0.000 (0.001)
Decrease	0.178*** (0.037)	-3.795*** (1.166)	-10.556** (4.832)	0.175 (0.358)	0.000 (0.001)
constant	0.141*** (0.022)	19.040*** (0.870)	60.916*** (3.399)	3.778*** (0.210)	0.971*** (0.001)
N	720	720	720	720	719
R2	0.082	0.044	0.016	0.003	0.001
p($Inc = -Dec$)	0.094	0.820	0.596	0.752	0.906
elast increase	-3.02	0.87	0.50	-0.45	0
elast decrease	-6.30	0.99	0.87	-0.23	0

Notes: OLS regressions. Robust standard errors in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%. The explanatory variables of interest are dummies indicating the *Increase* and *Decrease* group, respectively. The coefficients are relative to the omitted *Control* group. The outcome variables in columns (I) to (V) are: (I) *Ext. margin* is the extensive margin measured as a dummy variable indicating whether a subject quit the task immediately after seeing the treatment notification. (II) *Pics* is the number of images transcribed. (III) *Total time* is the time (in minutes) that a subject totally spent on working on the labor task. (IV) *Mean Time* is the time (in minutes) that a subject spent working on one image. (V) *Accuracy* is the share of characters that is transcribed correctly. *N* is the number of observations. *R2* is R-squared. p($Inc = -Dec$) is the p-value from a t-test testing whether the coefficients from the coefficients for the Increase and Decrease group add up to zero. *elast increase* and *elast decrease* are the elasticities in the treatment group that indicate how the respective outcome variable responds to the wage change, using the control group as the counterfactual (see section 4.3).

Appendix

A Distribution of pictures for all workers

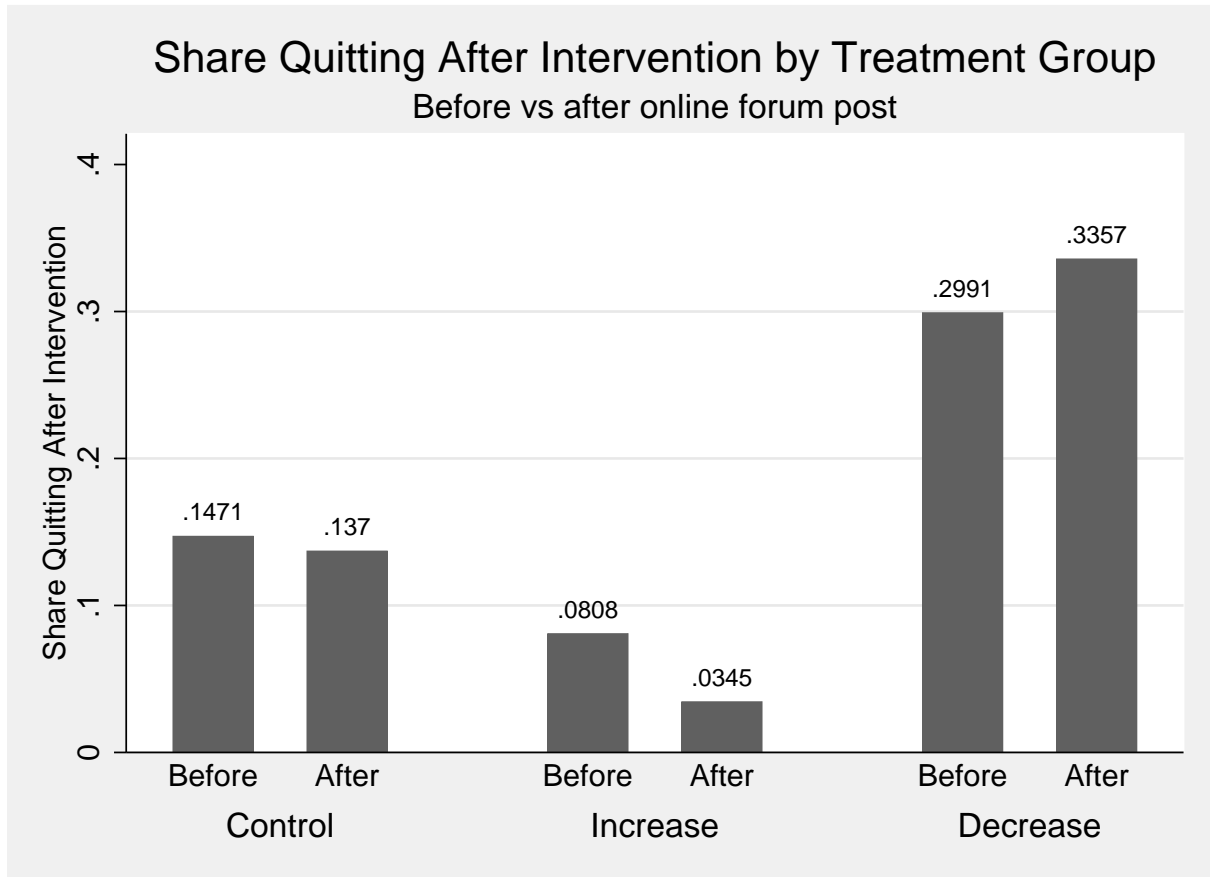
Figure 14: Histogramm of transcribed pictures



Notes: Histogramm of pictures described for all workers who worked on the task. The number of observations is 1152. Subjects saw the treatment notification after transcribing 6 pictures (indicated by the vertical line).

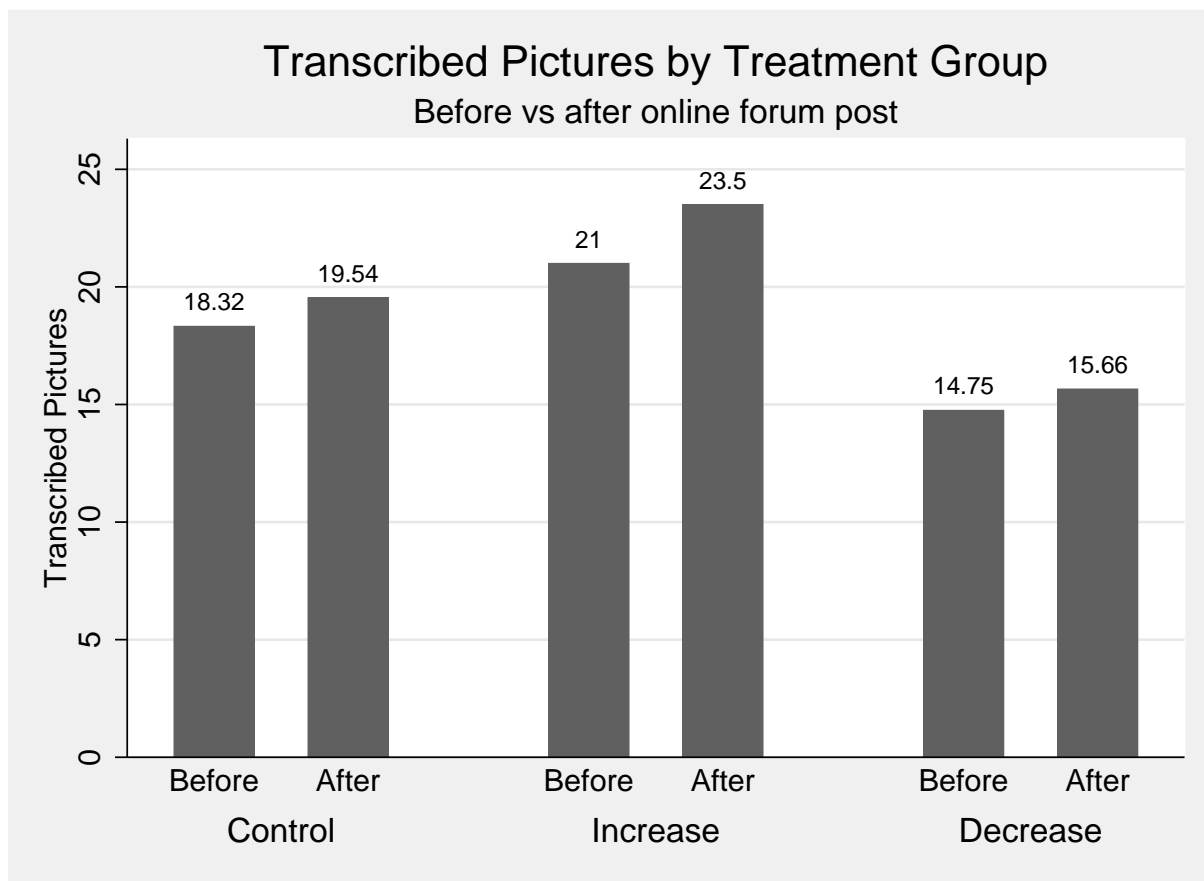
B Robustness: Effect of forum post

Figure 15: Extensive Margin. Before vs after forum post



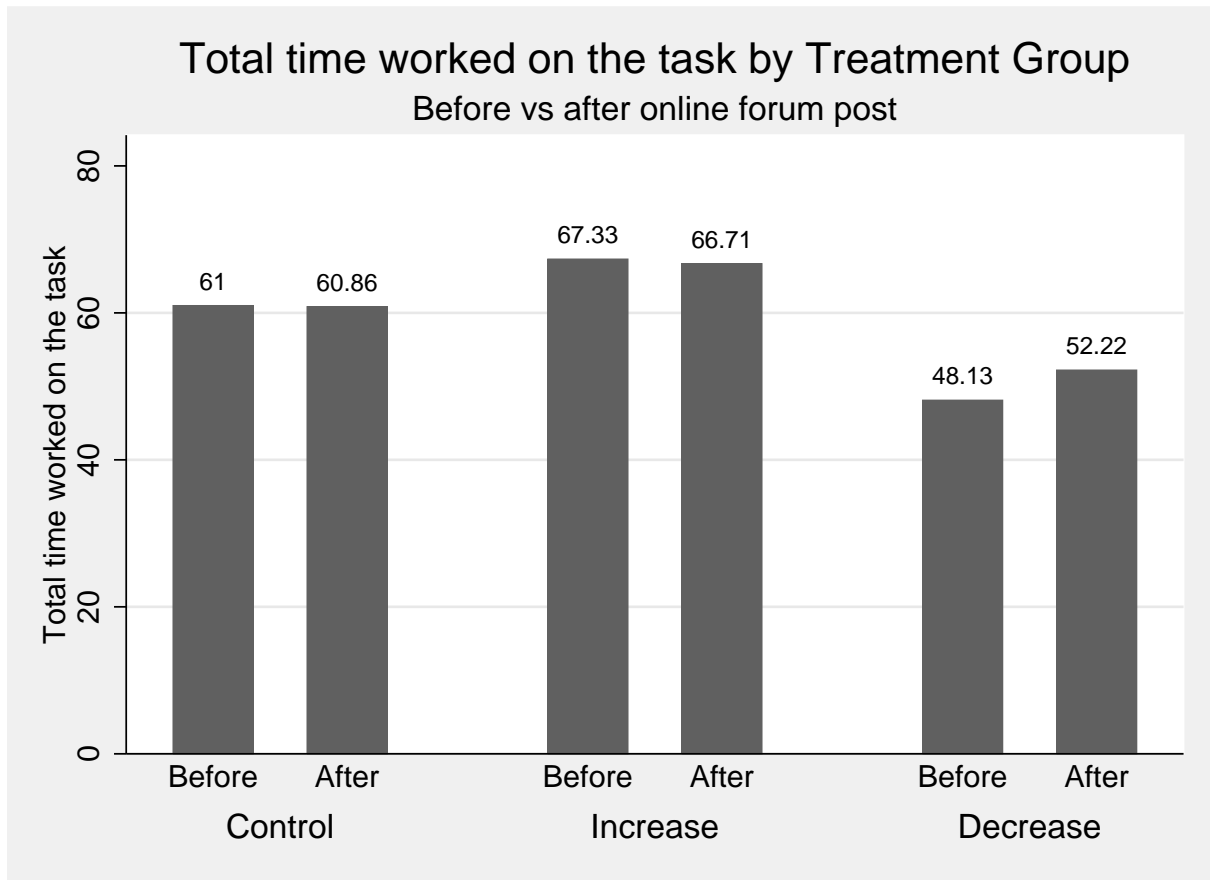
Notes: The Figure depicts the share of subjects in each group who quit the labor task immediately after seeing the treatment notification (i.e., share of subjects who transcribed six pictures but not a seventh one). *Before* and *After* indicate whether the observation was sampled before or after the task was discussed online (see section 4.2.) The number of observations sampled before the forum post is 318 with 102 subjects in the control group, 99 subjects in the "wage increase" group and 117 subjects in the "wage decrease" group. The number of observations sampled after the forum post is 402 with 146 subjects in the control group, 116 subjects in the "wage increase" group and 140 subjects in the "wage decrease" group. All subjects in the sample have transcribed at least six images.

Figure 16: Number of Transcribed Pics. Before vs after forum post



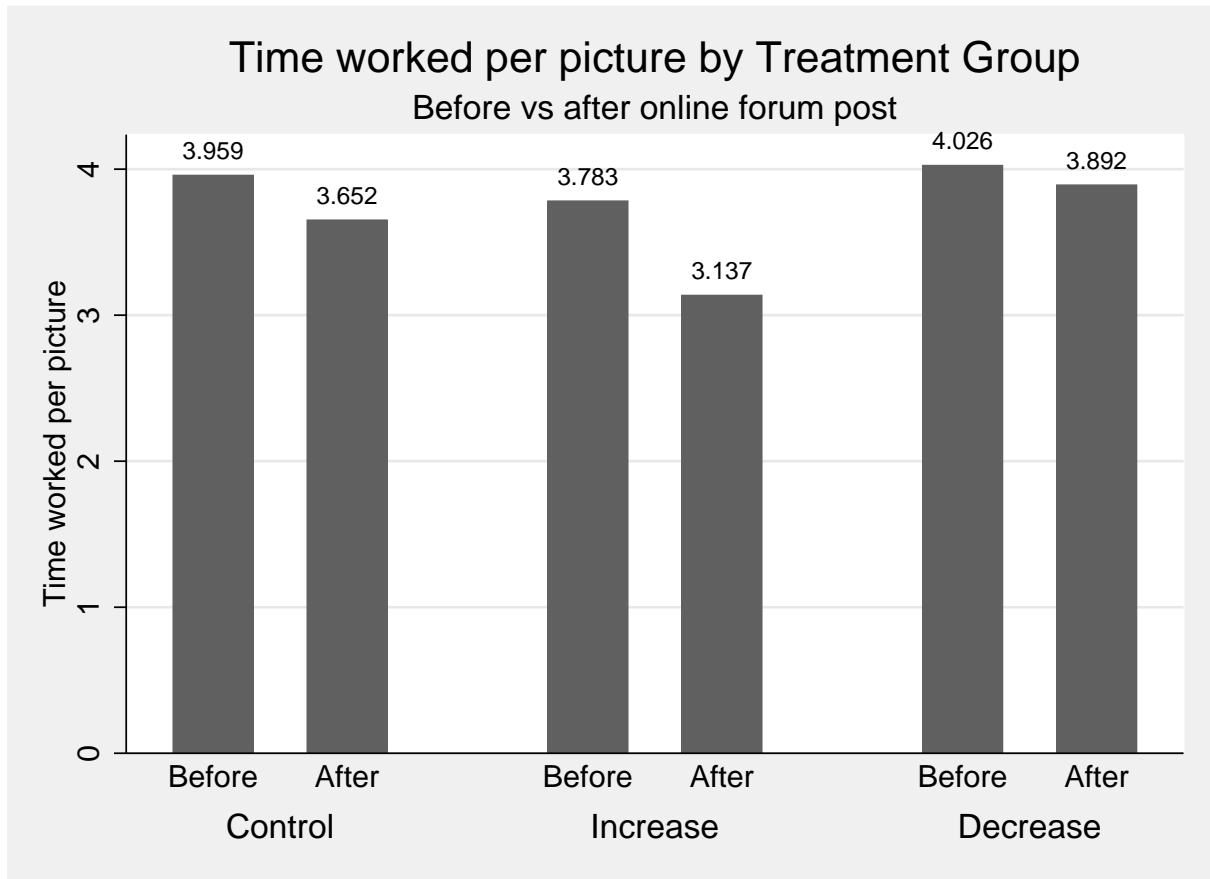
Notes: The Figure depicts for each group the average number of images that subjects transcribed. *Before* and *After* indicate whether the observation was sampled before or after the task was discussed online (see section 4.2.) The number of observations sampled before the forum post is 318 with 102 subjects in the control group, 99 subjects in the "wage increase" group and 117 subjects in the "wage decrease" group. The number of observations sampled after the forum post is 402 with 146 subjects in the control group, 116 subjects in the "wage increase" group and 140 subjects in the "wage decrease" group. All subjects in the sample have transcribed at least six images.

Figure 17: Total Time Worked. Before vs after forum post



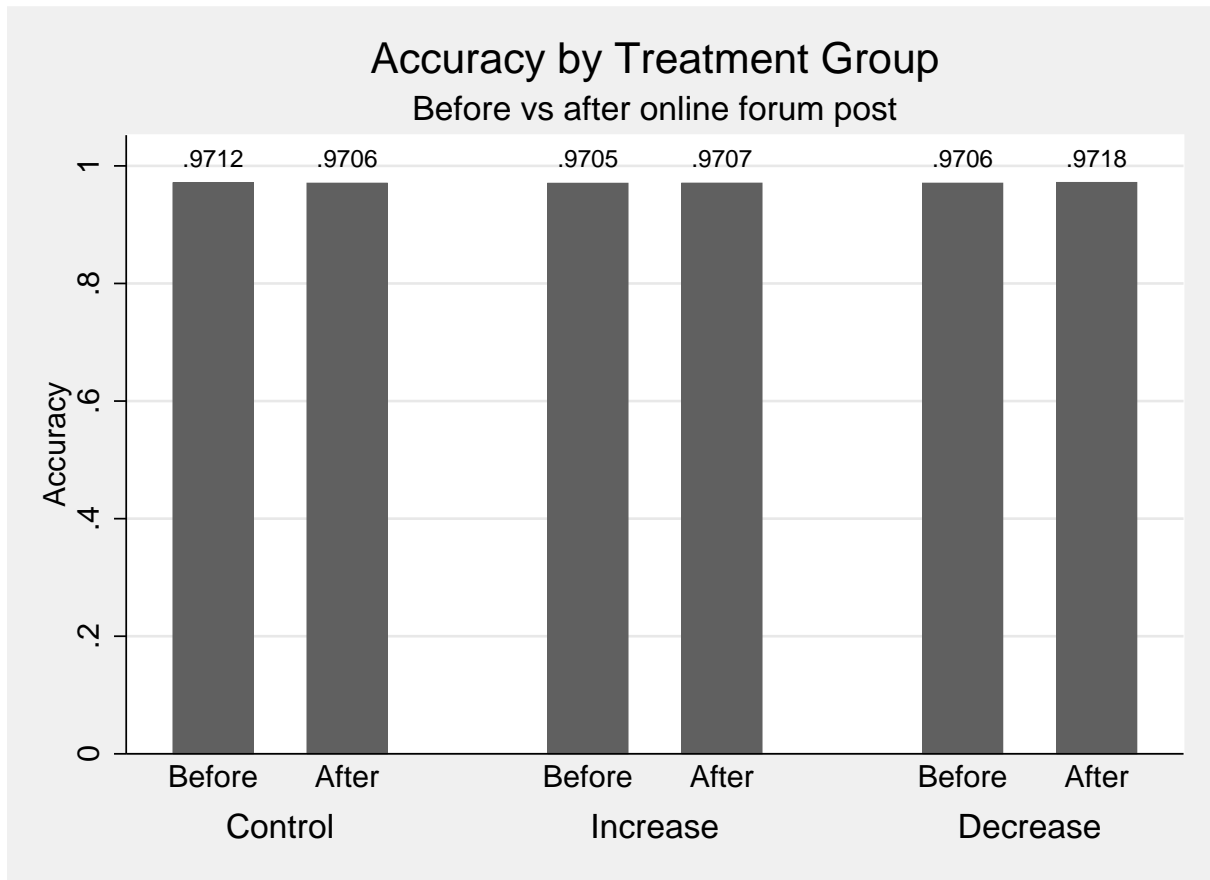
Notes: The Figure depicts for each group the average time (in minutes) that subjects totally spent on working on the labor task. *Before* and *After* indicate whether the observation was sampled before or after the task was discussed online (see section 4.2.) The number of observations sampled before the forum post is 318 with 102 subjects in the control group, 99 subjects in the "wage increase" group and 117 subjects in the "wage decrease" group. The number of observations sampled after the forum post is 402 with 146 subjects in the control group, 116 subjects in the "wage increase" group and 140 subjects in the "wage decrease" group. All subjects in the sample have transcribed at least six images.

Figure 18: Avg. Time per Hit. Before vs after forum post



Notes: The Figure depicts for each group the average time (in minutes) that subjects spent working on one image. *Before* and *After* indicate whether the observation was sampled before or after the task was discussed online (see section 4.2.) The number of observations sampled before the forum post is 318 with 102 subjects in the control group, 99 subjects in the "wage increase" group and 117 subjects in the "wage decrease" group. The number of observations sampled after the forum post is 402 with 146 subjects in the control group, 116 subjects in the "wage increase" group and 140 subjects in the "wage decrease" group. All subjects in the sample have transcribed at least six images.

Figure 19: Accuracy. Before vs after forum post



Notes: The Figure depicts for each group the average transcription accuracy, i.e., the average share of characters in each image that is transcribed correctly. *Before* and *After* indicate whether the observation was sampled before or after the task was discussed online (see section 4.2.) The number of observations sampled before the forum post is 318 with 102 subjects in the control group, 99 subjects in the "wage increase" group and 117 subjects in the "wage decrease" group. The number of observations sampled after the forum post is 402 with 146 subjects in the control group, 116 subjects in the "wage increase" group and 140 subjects in the "wage decrease" group. All subjects in the sample have transcribed at least six images.