CESifo WORKING PAPERS

7526
2019

February 2019

# Motivation and incentives in an online labor market

*Sebastian Fest, Ola Kvaløy, Petra Nieken, Anja Schöttner*

CESifo

## **Impressum:**

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

An electronic version of the paper may be downloaded
· from the SSRN website: www.SSRN.com
· from the RePEc website: www.RePEc.org
· from the CESifo website: www.CESifo-group.org/wp

CESifo *Working Paper No. 7526*
*Category 13: Behavioural Economics*

# Motivation and incentives in an online labor market

## Abstract

In this paper we present results from a large scale real effort experiment in an online labor market investigating the effect of performance pay and two common leadership techniques: Positive expectations and specific goals. We find that positive expectations have a significant negative effect on quantity - and no effect on quality - irrespective of how the workers are paid. On average, workers who receive positive expectations before they start to work, have a five percent lower output than those who do not. Goal-setting has no significant effect, neither on quantity nor quality. Performance pay, in contrast, has a strong positive effect on quantity, although we find no difference between high and low piece rates. Finally, we find no evidence of a multitask problem. Piece rates have no negative effects on the quality of work, even if it is fully possible for the workers to be less accurate and thereby substituting quality for higher quantity.

JEL-Codes: C930, M520, J330.

Keywords: non-monetary motivation, performance pay, field experiment.

*Sebastian Fest*
*University of Stavanger*
*UiS Business School*
*Norway – 4036 Stavanger*
*sebastian.fest@uis.no*

*Ola Kvaløy*
*University of Stavanger*
*UiS Business School*
*Norway – 4036 Stavanger*
*ola.kvaloy@uis.no*

*Petra Nieken*
*Karlsruhe Institute of Technology*
*Institute of Management*
*Kaiserstr. 89*
*Germany – 76133 Karlsruhe*
*petra.nieken@kit.edu*

*Anja Schöttner*
*HU Berlin*
*School of Business and Economics*
*Spandauer Straße 1*
*Germany - 10178 Berlin*
*anja.schoettner@hu-berlin.de*

February 8, 2019
We thank the participants of the Arne Ryde Workshop in Lund, Erasmus Workshop on Recognition and Feedback in Rotterdam, Stavanger Workshop on Incentives and Motivation, Seminar at Oslo Metropolitan University, Nordic Conference of Behavioral and Experimental Economics in Gothenburg for helpful comments and suggestions. Financial support from the Norwegian Research Council is gratefully acknowledged.

# 1 Introduction

An increasing number of workers do simple work in online labor markets. In the U.S., 24 million people receive parts of their income from online work, and the (world wide) annual growth of the so-called 'gig economy' is estimated to 14% (Kässi and Lehdonvirta, 2016). In contrast to employees within firms, online workers are governed by short term spot contracts where they engage with many different employers (Horton, 2010). They usually work from home, and do not have any personal contact with employers or colleagues, which imposes new challenges on firms to instill workforce motivation.

While it is typically straightforward to pay online workers for performance, motivating workers to work hard without paying them extra for doing so is more difficult. Within firms, workers are usually exposed to leaders who can motivate and inspire their workforce with words and actions. A large leadership literature argues that so-called transformational (or charismatic) leadership can improve performance and increase job satisfaction (for recent overviews, see Wang, Oh, Courtright, and Colbert, 2011; Robbins and Judge, 2013).

In contrast to transactional leaders who emphasize rewards in exchange for satisfying performance, transformational leaders inspire their followers with visions, positive expectations and challenging goals (Bass, 1985; Locke and Latham, 2002).[1] Indeed, controlled experiments have shown that motivating talks and charismatic leader behavior can have significant effects on workers' effort (Kvaløy, Nieken, and Schöttner, 2015; Antonakis, d'Adda, Weber, and Zehnder, 2014). However, such leadership instruments are harder to implement in online labor markets, where employers are typically left with digital messages that lack non-verbal elements such as visual or auditory clues which are main carriers of emotional communication. While face-to-face communication can evoke feelings of social presence and context, digital messages only offer a reduced toolbox of possible actions (Purvanova and Bono, 2009). It is still an open question if and how high quality leadership instruments work in the absence of personal contact.

In this paper we present results from a large scale real effort experiment on Amazon Mechanical Turk (MTurk), investigating the effect of two common leadership techniques: *Positive expectations* and *specific goals*. Communicating high performance expectations and expressing important purposes in simple ways is a key aspect of transformational leadership (e.g., Shamir, House, and Arthur, 1993; Tims, Bakker, and Xanthopoulou, 2011).

[1] There is also a growing literature on the economics of leadership, analyzing how motivational words and actions may improve performance (see Rotemberg and Saloner, 2000; Van den Steen, 2005; Dur, Non, and Roelfsema, 2010; Kvaløy and Schöttner, 2015; Hermalin, 2015).

A large literature has both theorized and documented that leaders who are able to articulate positive expectations and specific goals succeed in motivating their workforce (e.g., Judge and Bono, 2000; Locke and Latham, 2002). However, these leadership techniques have not been studied in online labor markets. MTurk is a major online labor market with more than half a million registered workers world wide. We recruited experienced U.S. workers to work on a short transcription task, enabling us to study the effect of expectations, goals, and performance pay on both the quantity and the quality of work.

In the expectations treatments and the goal treatments, we use digital messages to motivate workers before they start working. In particular, we inform workers on screen, in the expectations treatments, that we are happy that they will work for us, and that we know they are diligent workers with an impressive reputation. In the goal treatments, we ask them to achieve a quite ambitious output level within the working period. In the performance pay treatments, we vary between no piece rate, a low piece rate and a high piece rate. The workers were explicitly paid for quantity, although we could also measure the quality of their performance (in terms of correct transcriptions).

According to standard economic theory, we should only expect to find effects on quantity in the performance pay treatments. If it is possible to substitute quality for quantity, piece rates should also reduce the quality of work, which is known as the multitask problem (e.g., Holmström and Milgrom, 1991; Baker, 1992). As argued above, leadership theories, on the other hand, predict that non-monetary motivation in terms of positive expectations and well-designed goals can enhance performance even in the absence of performance pay. However, it is an open question as to whether these instruments have an impact in an anonymous online spot labor market.

We are also interested in the interaction effects between performance pay and non-monetary motivation. Are those instruments substitutes or complements if they are employed together? While the latter is suggested by the results of Kvaløy et al. (2015), it is unclear if such a complementarity can be triggered with simple digital messages.

Our main results are as follows: We find that positive expectations have a significant negative effect ($p < 0.01$) on quantity – and no effect on quality – irrespective of how the workers are paid. On average, workers who receive positive expectations before they start to work have a five percent lower output than those who do not. Goal-setting has no significant effect, neither on quantity nor quality. Performance pay, however, has a strong positive effect on quantity, although we find no difference between a low and high piece rate. Furthermore, we find no interaction effects of varying piece rates and setting expectations. Expressing output goals, on the other hand,

renders monetary incentives ineffective for increasing quantity. Finally, we find no evidence of a multitask problem. Rather, we observe a slightly positive relationship between quantity and quality across all treatments, including performance pay treatments that explicitly paid for quantity.

Our results shed light onto the question of how workers are motivated in online labor markets. Trying to implement simple leadership techniques in short term impersonal interactions may not only be useless, it may in fact be detrimental to effort. This does not imply that any attempt to motivate online workers with other instruments than money does not pay off. For instance, it has been shown that task significance and social comparisons can improve performance in online markets (Chandler and Kapelner, 2013; DellaVigna and Pope, 2017). However, our experiment demonstrates that online employers cannot simply adopt well-established leadership tools developed in more traditional organizational contexts. A possible explanation is that online workers do not expect online employers to behave as if they are traditional long-term employers who are "happy that the workers will work for them." Positive expectations, reminders of achievements, or specific goals may be perceived as non-credible and potentially provocative (Farson, 1963). Further research is needed in order to test this hypothesis.

Our experiment also shows that the introduction of very small piece rates works surprisingly well in the context we consider, while the marginal effect of increasing the level of monetary incentives is close to zero. This result contrasts with Gneezy and Rustichini's (2000) "Pay enough or don't pay at all" result, and is more in line with DellaVigna and Pope (2017) and Pokorny (2008), who find a strong effect of introducing a small piece rate, but, respectively, a low or even negative effect of increasing the piece rate. Interestingly, we find no negative effects of piece rates on the quality of work, even if it was fully possible for the workers to be less accurate, thereby substituting quality for higher quantity.

Indeed, the empirical evidence regarding the relevance of multitasking problems in practice is mixed. Hong, Hossain, List, and Tanaka (2013) present a field study on Chinese factory workers that is in strong support of the multitasking theory. The authors argue that the key distinction of their setting relative to many others (that are not in line with the multitasking theory) is that quality is not only unrewarded but also truly unobservable by the principal, which is crucial to fully eliminate reputational concerns of workers. In a similar spirit, Al-Ubaydli, Andersen, Gneezy, and List (2015) propose that agents' uncertainty about the principal's monitoring technology can even lead to higher quality under piece rates than under fixed wages. We conducted clarification treatments to test whether the absence of a quality-quantity trade off in our setting is driven by asymmetric information concerning the implications of low-quality work. We find that this is not the case. Our results may thus indicate that online workers

put some pride in doing a decent job, and are not driven by monetary or reputational incentives alone.

In simple work settings such as ours, specific and challenging (yet attainable) goals are considered to be effective means of motivation in the fields of psychology and management (e.g., Locke and Latham, 1984, 2002) and economics (e.g., Goerg and Kube, 2012; Corgnet, Gómez-Miñambres, and Hernán-Gonzalez, 2015, 2018). Economic research suggests that goals motivate workers because they serve as reference points and thus influence workers' decisions when their utility is reference-dependent (Corgnet et al., 2015, 2018). Interestingly, DellaVigna and Pope (2017) do not find statistically significant support for workers on MTurk exhibiting reference-dependent utility, which may explain why workers do not respond to goals in our study.

On a more general level, our study is related to an increasing number of recent papers that utilize online labor markets to study work incentives or participation decisions (e.g., Chandler and Kapelner, 2013; DellaVigna and Pope, 2017; de Quidt, 2017; List and Momeni, 2017). However, none of these papers examines the effectiveness of traditional leadership instruments in combination with monetary incentives. In contrast to DellaVigna and Pope (2017), our task has both a quantity and a quality dimension, and we study the interaction of monetary incentives and non-monetary leadership techniques.

The remainder of the paper is organized as follows. We explain the design, our hypothesis and experimental procedures in detail in Section 2 and present the results in Section 3. The paper closes with a discussion in Section 4.[2]

# 2 The experiment

In the following, we discuss the treatments and describe the experimental set-up in detail. Using a factorial design, we employ nine between-subject treatments for our experiment. This design allows us to study the impact of leadership techniques and performance pay on worker performance as well as the interaction of these two features on a large sample of online labor market participants.

## 2.1 Design

To study behavior in an online labor market, we chose to conduct our experiment on Amazon Mechanical Turk, one of the most prominent and

[2]Data and code to reproduce all estimates are available at `https://github.com/sebfest/motivation_and_incentives`

widely used platforms that currently exists (Peer, Brandimarte, Samat, and Acquisti, 2017). MTurk offers firms the opportunity to outsource small, manual tasks to a large number of online workers. Potential employers post job offers on the MTurk platform and can specify a set of criteria that workers have to meet in order to be allowed to work on the task. These screening options can either be related to the reputation of the worker, such as the total number of tasks the worker has previously completed, the share of tasks that the worker previously got approved, or to specific demographics of the worker, such as location, age, or gender.

Workers who are registered on the MTurk platform can browse among available tasks that fit their criteria or search for job offerings posted by particular employers or according to keywords used in the task description. This description typically contains information about the offered payment as well as the task duration. Workers who accept a work task then have to complete the task within a specified time interval set by the employer. After task completion, the employer reviews the submitted task and can approve and pay the worker or reject the work if necessary. In the case of a rejection, the approval rate of the worker drops, leading to a loss of the worker's future potential to find suitable job offers.[3]

In order to measure any effects on work performance in the experiment with respect to quantity and quality, we chose a text transcription task. In particular, we asked workers to type text from a series of fragments taken from an ancient Latin text for a total duration of 10 minutes. The fragments had an average length of about 50 characters and were shown as a picture on the screen, such that workers were prevented from simply copying and pasting the text. Workers only saw a single fragment at a time and had to submit their transcription in order to receive a new fragment on their screen. The typesetting of the letters for all fragments was historic so that some letters were harder to read. The task therefore requires effort, attention, and diligence.

Using the transcription task, we employed three main monetary incentive treatments called *No piece rate*, *Low piece rate*, and *High piece rate* to investigate reactions to variations in the payment structure.[4] While workers received no extra pay for the number of submitted fragments in the *No piece rate* treatment, workers received a piece rate of \$0.01 per submitted fragment in the *Low piece rate* treatment. In the *High piece rate* treatment, we increased the piece rate to \$0.05. We informed workers about the piece rate in the following way: *"In addition, you will receive a bonus of \$0.01 (\$0.05) for each completed fragment. The compensation will be sent to you within two days after the completion of this HIT."* The increase in the piece

[3]An approval rate of 98% is often deemed critical in this regard among workers and employers.

[4]The instructions for the experiment can be found in section 5.2 of the Appendix.

rate of 5 Cents leads to a considerable potential earnings increase for the ten minute task. A worker who submits 30 fragments, for example, yields a $1.2 higher payment in the *High piece rate* treatment than in the *Low piece rate* treatment.

To investigate whether up-front motivation, in particular expectation setting from the employer or the expression of output goals, have an impact on work performance, we conducted two main non-monetary incentive treatments called *Expectation* and *Goal*. In both treatments, workers saw a simple screen before starting to work on the task. In the *Expectation* treatment, workers read: *"Before you start, we want to emphasize how happy we are that you've decided to work for us. You've proven to be a successful and diligent worker on MTurk with an impressive approval rate!"* In the *goal* treatment, workers read: *"Efficient work is important. Please try to submit at least 25 fragments."* Workers could leave the message screen at any time by clicking on a button to proceed to the work task. In order to check for the interaction effect of monetary and non-monetary incentives on work performance, we combine the *Expectation* and *Goal* treatments with each piece rate payment scheme, respectively. The resulting 3x3 treatment design is summarized in Table 1.[5]

[Table 1 about here]

## 2.2 Sample and procedures

For our experiment, we invited a total of 2700 workers from Mturk from the fall of 2016 to spring 2017. Workers responded to a job posting offering a ten minute work task for a $2 payment that had to be completed within one hour. Our selection criteria for workers stipulated that subjects on Mturk needed to have a total number of 500 previously approved tasks and a task approval rate of 98 percent. In addition, only workers who indicated their location as the United States were eligible for participation.[6]

Workers who accepted the job offer followed a link to an external website that we used for data collection. After workers gave their consent to

[5] As a robustness check, we conducted treatments where we only cut off the concern for the approval rate by stating that the work would be approved automatically. As the results do not differ compared to the other treatments, we pooled the data with the respective treatments.

[6] For the design and conduct of the experiment, we closely followed guidelines mentioned in a series of articles that discuss the use of Mturk in behavioral research (Paolacci, Chandler, and Ipeirotis, 2010; Horton, Rand, and Zeckhauser, 2011; Berinsky, Huber, and Lenz, 2012; Mason and Suri, 2012; Crump, McDonnell, and Gureckis, 2013; Paolacci and Chandler, 2014). This includes that measures were taken for excluding duplicate workers, workers who participated in earlier related experiments, and checking for workers who attempt to self-select into treatment. We find that 28 workers in our sample restart their work task. This does not result in any selection effect.

participate in the study and finished reading the task instructions, they started working on the task. The task stopped automatically after ten minutes. At the end, all workers answered a short survey and received a code for verification.[7]

[Table 2 about here]

The survey contained demographic questions as well as questions regarding the worker's familiarity with Latin and the device used to complete the task. Table 2 provides an overview of the background characteristics of subjects participating in the experiment. Workers are, on average, 36 years old, possess a two year college degree, and are only vaguely familiar with Latin. About five percent use a mobile device to complete the task. The sample also contains an equal number of male and female workers. Importantly, we observe that the treatments were balanced with respect to all of these characteristics.

Altogether, workers spend on average 13 minutes to complete the experiment. Average payments made amounted to $2.80, including the $2 participation fee. All payments were made electronically. Participation fees were paid out soon after the experiment. Payments based on worker performance were transferred within two days after the study was conducted.

## 2.3 Hypotheses

Following standard economic theory, we should observe increased output levels in the performance pay treatments compared to the fixed wage case: Workers exert effort to the extent that marginal effort costs equal marginal monetary gains. Hence, fixed pay should make workers exert effort at a convenient level, while piece rates should make them work harder. In contrast to many other studies (e.g., DellaVigna and Pope, 2017), our task has a quantity and a quality dimension. Therefore, workers in the piece rate treatments face a multitasking problem. If they want to maximize their payment, they have to type faster which could result in more errors in the submitted output. We expect that workers deliver lower quality in the piece rate treatments compared to the no piece rate treatment, and a negative correlation between quantity and quality in those treatments.

Following the leadership theories discussed in the introduction, we should also expect goals and positive expectations to evoke higher performance for a given fixed wage. However, leadership scholars emphasize the importance

[7]Four workers accepted the invitation but never actually worked on the task and are therefore missing from the sample. In addition, the timer of the work task did not work properly for 16 workers who had to be excluded after data collection.

of enduring attention and devotion from leaders in order to improve workers' performance (Robbins and Judge, 2013). Leaders should also combine goals and expectations with organizational identity and commitment (Basu and Green, 1997; Walumbwa, Avolio, and Zhu, 2008). But these latter leadership ingredients are not easy to implement in online labor markets. Using simple leadership techniques to motivate online workers who work for very short periods, therefore, may not have the desired effects.

We are also interested in the interaction effects between performance pay and non-monetary motivation. Psychological theories of motivation predict that monetary incentives alone can crowd out intrinsic motivation and thereby weaken performance (e.g., Deci, 1971; Deci and Ryan, 1971). However, recent behavioral economics theories (e.g., Bénabou and Tirole, 2003; Ellingsen and Johannesson, 2008) imply that crowding out effects may be reduced or eliminated if the principal can resolve informational asymmetries. For example, motivational talk by a leader can help clarify the nature of the task or the characteristics of the principal. Indeed, Kvaløy et al. (2015) find in a field experiment that motivational talk (including positive expectations) enhances the effectiveness of performance pay. In line with this, we hypothesize that expectations and performance pay are complements also for the online workers we study.

# 3 Results

## 3.1 Quantity

In the following analysis, we address the issue of whether changes in monetary as well as non-monetary incentives affect workers' productivity in the text transcription task. We answer this question by first focusing on the average number of fragments submitted in each treatment. In particular, we test for an effect of increasing the piece rate per submitted fragment, test for the effect of using different up-front motivational messages on worker output, and also test for an interaction between these two dimensions on the productivity of workers. We do so by using ordinary least squares (OLS) regressions with robust standard errors.

[Table 3 about here]

Table 3 presents results from a series of linear regressions, where the dependent variable in each regression captures the number of fragments submitted per worker. The first column reports main effect estimates for increasing the piece rate from zero in the *No piece rate* treatments to \$0.01 and \$0.05 in the *Low piece rate* and *High piece rate* treatments, respectively. We find that both changes yield a significant positive effect on

worker productivity. In particular, the estimate results reveal an increase in average output of 1.39 fragments ($p < 0.001$) for a low piece rate and an increase in average output of 1.41 fragments ($p < 0.001$) for a high piece rate. Relative to the *No piece rate* treatments in which workers submit 22.2 fragments on average, these changes correspond to a relative increase in worker productivity of about 6.3 percent for both piece rates, respectively. Interestingly, we find that the motivational effect of changing the monetary incentives does not depend on the size of the incentive change itself. Specifically, while we estimate a positive effect of both piece rates on productivity, we cannot identify any difference between the two piece rate treatments ($diff = 0.028,\ p = 0.948$). Thus, even offering the minimum piece rate payment of one Cent increases worker output as much as a five times higher piece rate.

The second column in Table 3 lists main effect estimates for the two up-front motivational messages. We find that while communicating specific output goals to workers up-front has no effect on the number of fragments submitted, conveying positive expectations about the worker prior to work lowers productivity in the work task.[8] More precisely, we estimate that the expression of output goals insignificantly lowers workers' productivity by 0.3 fragments ($p = 0.425$). In contrast, when we set positive expectations towards workers rather than using no up-front motivational text at all, the average number of submitted fragments substantially decreases by 1.2 fragments ($p = 0.006$). Relative to the *Neutral* treatments, where workers submit an average of 23.7 fragments, this decrease is equivalent to a five percent drop in productivity. Moreover, we also identify that the negative effect of setting positive expectations is significantly stronger than articulating specific output goals to workers ($diff = 0.873,\ p = 0.033$). Overall, this suggests that sending a simple motivational message before the working phase either seems to have no effect on motivation at all or even impairs workers' motivation, leading to a drop in work output.[9]

Column three in Table 3 reports estimate results from a fully saturated regression specification that includes interaction terms for each piece rate combined with the introduction of the two different up-front motivational messages. This specification allows us to test for the existence

[8]For both treatments that employ non-monetary motivational techniques, we present workers a screen with an up-front motivational text prior to work. Figure S1 in the Appendix shows that workers spend on average approximately 6 and 16 seconds reading the motivational texts in the Goal and Expectation treatment, respectively. Note that our goal message is substantially shorter than our expectation message.

[9]Although we cannot identify any effect of setting a performance goal on the average number of fragments submitted, we do find that the goal setting harmonizes the exerted effort levels of workers. Specifically, the variance in produced output is significantly lower in the *Goal* treatments than in the *Neutral* and *Expectation* treatments (two-sided Levene's variance comparison test, $p < 0.001$ for both comparisons, respectively)

of complementary and substitutional relationships of monetary and non-monetary motivational techniques. While we find no significant interaction between setting positive expectations towards workers prior to work and increased monetary incentives, we identify that the expression of explicit output goals to workers curbs the positive effects that result from increasing the monetary reward per submitted fragment. This can be seen by the set of treatment interaction variables which measure the difference-in-difference effect of introducing a low and high piece rate with our expectation and goal setting, respectively. Specifically, while the low and high piece rate significantly increase the average number of submitted fragments ($p = 0.012$ *and* $p = 0.003$, *respectively*), the *Low piece rate* $\times$ *Expectations* and *High piece rate* $\times$ *Expectations* indicator variable estimates remain insignificantly small ($p = 0.801$ *and* $p = 0.883$, *respectively*), suggesting that conveying positive expectations to the worker up-front does not render monetary incentives ineffective for increasing workers' motivation. On the other hand, the magnitudes of the *Low piece rate* $\times$ *Goal* and *High piece rate* $\times$ *Goal* indicator variable estimates indicate that the disclosure of specific output goals to workers prior to work offsets the positive impact on worker productivity that result from increased monetary incentives ($p = 0.176$ *and* $p = 0.031$, *respectively*).[10]

Regression results presented in column four of Table 3 add a set of worker background variables to the regression specification used in column three. The set includes variables for gender, age, education, device used for the work task and knowledge of Latin. From this set of background variables, we find that older workers submit, on average, fewer fragments while more educated workers and women show a higher work productivity in the task. Moreover, the knowledge of Latin is not predictive for worker output in the text transcription task whereas mobile users, on average, submit five fragments fewer than non-mobile device users.

## 3.2 Quality

In contrast to many other studies, we can assess work performance not only through quantity but also through the quality of the submitted fragments. In this section, we therefore check whether changes in monetary as well as non-monetary incentives affect the quality of the delivered output. Specifically, we test for an effect of increasing the piece rate per submitted fragment as well as for an effect of using different up-front motivational messages on the quality of the submitted fragments, and also test for an

[10]While we cannot conclude that the *Low piece rate* $\times$ *Goal* estimate is statistically significant here, we identify a significant effect for the interaction term when the regression specification includes a series of control variables ($p = 0.090$) as presented in column four in Table 3.

interaction between these two dimensions on the quality of the submitted fragments.

In order to analyze the quality of the work, we construct an error score as follows. First, for each submitted fragment, we calculate the Levenshtein edit distance (Levenshtein, 1966), i.e., we compute the minimum number of edit operations involving the insertion, deletion, or substitution of individual characters which are required to transform the submitted fragment into the correct fragment. Specifically, we apply a unit cost to each edit operation and allow for the fact that workers could use the "?" character as a wildcard if they were unable to identify the actual character in the presented fragment. As shown in equation (1), we then normalize the processed edit distance by the upper bound of transforming the submitted fragment into the correct fragment, i.e., we divide the processed edit distance by the length of the longer string, obtaining a ratio of dis-similarity of the two fragments that we interpret as the error score. Formally we have:

$$\text{Error score} = \frac{\text{Edit distance(answer, solution)}}{\max(\text{len(answer)}, \text{len(solution)})} \in [0, 1] \tag{1}$$

Table 4 presents results from a series of random-effects panel regressions, where the dependent variable in each regression represents the error score of a fragment submitted by a worker.[11] The first and second column reports main effect estimate results of changing the monetary and non-monetary incentives, respectively. We find that neither changes in the piece rate nor differences in the up-front motivational messages have any effect on the quality of work. More precisely, we find that workers submit, on average, over time and across treatments, fragments that have an error score of about 0.018, i.e., fragments which have a dis-similarity of about 1.8 percent with the correct fragment.

Column three of Table 4 contains estimate results from a fully saturated regression specification that includes interaction terms for both treatment dimensions. The estimate results corroborate the overall impression from before and reveal that the quality of fragments does not systematically vary across treatments. Estimate results presented in the fourth column of Table 3 include controls from a set of background variables including gender, age, education, device used for the work task, and knowledge of Latin. From all background variables, we find that female workers as well as more educated workers submit, on average, fragments with a smaller error score while mobile users deliver fragments that are more error prone.

[Table 4 about here]

[11] A Breusch-Pagan Lagrange multiplier test consistently rejects the null hypothesis of no significant difference across units for each specification. We therefore estimate treatment effects using a random-effects model.

### 3.3 Quantity vs. Quality

Workers in our experiment could either type very fast and submit a large number of fragments in the work task, or, they could put increased care into the correctness of their submitted fragments, which would result in a smaller number of submitted fragments. In this section, we test for the existence of this multi-tasking problem and check whether the trade-off between quantity and quality varies according to the monetary or non-monetary incentives that we give to workers.

[Figure 1 about here]

Figure 1 plots the number of submitted fragments against the timed average error score for all submitted fragments for each worker by treatment. Across all treatments, we find no indication for the existence of a multitasking problem. Specifically, from the set of sample correlation coefficients that we estimate for each treatment cell, we cannot identify any single significant negative linear relationship between the number of submitted fragments and their average quality. In marked contrast to our initial hypothesis, we consistently find across all treatments that workers who submit a larger number of fragments also submit fragments that are characterized by a lower average error score.[12]

Based on the above results and the results from Section 3.1 and Section 3.2, we therefore infer the following. First, the effect of monetary incentives on the number of submitted fragments does not come at the expense of quality. Across all *Low piece rate* and *High piece rate* treatments where we found a relative increase in output of more than six percent, the correlation between the number of submitted fragments and the average error score is significantly below zero ($r = -0.15$ and $r = -0.16$, respectively, both $p < 0.001$). Second, conveying positive expectations about the worker prior to work leads workers to work more slowly on the task but not more accurately. Specifically, across all *Expectation* treatments for which we estimated a five percent lower number of submitted fragments, the correlation between the number of submitted fragments and the average error score is significantly negative ($r = -0.12,\ p < 0.001$).

# 4 Discussion

Sending a simple message before the work phase inhibits or even decreases motivation of workers. While this result is puzzling at first sight, it may

[12]Table S1 in the Appendix shows regression results for regressing the averaged error scores per worker on the number of submitted fragments per worker. We allow for intercepts and slope parameters to vary separately as well as in combination. We identify no significant differences in slope or intercept parameters across treatments.

show that non-monetary motivational interventions can also have negative performance effects. However, the reduction in performance could also simply be due to the interruption before the working stage itself and not due to the content of the message.

If the drop in performance is simply due to the interruption itself and not the content of the message, we would expect workers in the *Goal* treatments to also react negatively to the message since they spend a substantial amount of time reading the performance goal message as well (See Figure S1 in the Appendix). However, we find no indication of a negative effect of our goal message. Hence, we do not believe that the negative effect of expectation on productivity is driven by interrupting workers per se but by the content of the message.

Another unexpected outcome of our experiment is that we do not find any evidence for a multitasking problem. This, despite the fact that workers in the piece rate treatments have a high incentive to trade-off quality for quantity. A possible explanation for this result is that workers were concerned about not receiving their piece rate payment if the delivered quality was too low and therefore, in response, worked more slowly on the task.

To address the above issue, we employed additional *Clarification* treatments where we explicitly informed workers that we would not check the quality of their submitted fragments. In particular, we implemented a special emphasis on the security of the piece rate payment regardless of whether the fragment was correct or not by stating to workers that *"In order to pay the bonus in due time, we pay it for submitted fragments without controlling for typing errors. Once you have completed the HIT, you will be approved automatically, which means that your performance will not affect your approval rate".* In the *Clarification* treatments, there was no need for workers to work diligently on the task in order to avoid being rejected and not receive the piece rate.

Using this clarification, we employed four additional treatments on a sample of 400 workers, including two treatments with a low and high piece rate payment scheme without any expectation setting and two treatments with the low and high piece rate payment scheme in combination with the setting of expectations prior to work.[13] If the concerns about receiving work payment affected how workers in the original treatments evaluate the multitasking problem, we would expect to find a change in how workers trade-off quality for quantity when we signal that we do not control for mistakes.

[Figure 2 about here]

[13]Two workers accepted the invitation but never actually worked on the task. In addition, two other workers had to be excluded after data collection because their timer did not function properly.

Figure 2 plots the number of submitted fragments against the timed average error score for all submitted fragments for each worker by treatment for the additional sample. With the additional clarification regarding the absence of quality control, we still find no evidence that workers who submit a larger number of fragments also submit fragments of lower quality. Specifically, none of the sample correlations coefficients for the additional set of treatments is significantly larger than zero. Moreover, across all new clarification treatments, we estimate a sample correlation of $r = -.14,\ (p < 0.001)$ between the number of submitted fragments and the average error score. Strikingly, we obtain a similar coefficient for the treatment counterparts in the original experiment where we did not use any additional clarifying statement ($z = 0.035,\ p = 0.972$). This suggests that the absence of a quality-quantity trade off in our original setting is not driven by asymmetric information concerning the implications of low quality work.[14]

# 5 Conclusion

In contrast to employees within traditional firms, workers in online labor markets are not exposed to leaders who can inspire them with words and actions. This makes motivation more challenging. In this paper we have presented results from a large scale experiment on Amazon Mechanical Turk, investigating the effect of performance pay and two common leadership techniques: Positive expectations and specific goals. We study how online workers' productivity is affected both by up-front motivational messages and monetary incentives.

Whereas monetary incentives enhance output, the motivational messages have either no effect, or a negative effect on the workers' performance. In particular, we find that expressing positive expectations has a significant *negative* effect on quantity – and no effect on quality – irrespective of how the workers are paid. Goal-setting has no significant effect, neither on quantity nor quality. Performance pay has a strong positive effect on quantity, although we find no difference between a low and high piece rate. Interestingly, we find that the quality of work varies independently of monetary and non-monetary incentives. Moreover, we find no evidence of a multitask problem. Rather, we observe a slightly positive relationship

[14]In Table S2 and Table S3 in the Appendix, we provide regressions of quality on quantity, estimating slopes and intercept parameters for each additional treatment as well as parameters comparing the overall quantity quality trade-off with and without the additional clarification statement, respectively. We find no difference in the overall trade-off. In addition, we also present regressions of quantity and quality on a set of treatment variables in Table S4 and Table S5. We find no effect of the clarification statement on quantity or quality.

between quantity and quality across all treatments, including performance pay treatments that explicitly paid for quantity.

Our experiment demonstrates that online employers cannot simply adopt well-established leadership tools developed in more traditional organizational contexts. A possible explanation is that online workers do not expect online employers to behave as if they are traditional long-term employers who are "happy that the workers will work for them." Positive expectations, reminders of achievements, or specific goals may be perceived as non-credible and potentially provocative (Farson, 1963). However, more research is needed in order to understand when motivational messages may actually lower performance. Psychologists have studied praise as a social reinforcer and found that praising people can be ineffective or even dysfunctional (Delin and Baumeister, 1994). For example, whether praise enhances or undermines children's intrinsic motivation strongly depends on environmental and individual characteristics (e.g., Henderlong and Lepper, 2002). Moreover, recent research in marketing shows that praising messages in written form can change people's behavior, but are more effective when coupled with an assertive tone (Grinstein and Kronrod, 2016).

As a final remark, an important takeaway from our study is that online workers seemingly put pride in doing a decent job irrespective of extrinsic motivation. The workers' productivity is remarkably stable across all treatments. Even when we remove all monetary and non-monetary incentives to exert any effort, they still work hard.

# References


Al-Ubaydli, Omar, Steffen Andersen, Uri Gneezy, and John A List (2015). "Carrots that look like sticks: Toward an understanding of multitasking incentive schemes," *Southern Economic Journal*, 81(3): 538–561.

Antonakis, John, Giovanna d'Adda, Roberto Weber, and Christian Zehnder (2014). "Just words? Just speeches? On the economic value of charismatic leadership," *working paper*.

Baker, George P. (1992). "Incentive contracts and performance measurement," *Journal of Political Economy*, 100(3): 598–614.

Bass, B. M. (1985). *Leadership and Performance Beyond Expectations*, Free Press.

Basu, Raja and Stephen G Green (1997). "Leader-member exchange and transformational leadership: An empirical examination of innovative behaviors in leader-member dyads," *Journal of Applied Social Psychology*, 27(6): 477–499.

Bénabou, Roland and Jean Tirole (2003). "Intrinsic and extrinsic motivation," *Review of Economic Studies*, 70: 489–520.

Berinsky, Adam J., Gregory A. Huber, and Gabriel S. Lenz (2012). "Evaluating online labor markets for experimental research: Amazon.com's Mechanical Turk," *Political Analysis*, 20(3): 351–368.

Chandler, Dana and Adam Kapelner (2013). "Breaking monotony with meaning: Motivation in crowdsourcing markets," *Journal of Economic Behavior & Organization*, 90: 123–133.

Corgnet, Brice, Joaquín Gómez-Miñambres, and Roberto Hernán-Gonzalez (2015). "Goal setting and monetary incentives: When large stakes are not enough," *Management Science*, 61(12): 2926–2944.

Corgnet, Brice, Joaquín Gómez-Miñambres, and Roberto Hernán-Gonzalez (2018). "Goal setting in the principal-agent model: Weak incentives for strong performance," *Games and Economic Behavior*, 109: 311–26.

Crump, Matthew J. C., John V. McDonnell, and Todd M. Gureckis (2013). "Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research," *PLoS ONE*, 8(3): e57410.

Deci, E. L. (1971). "Effects of externally mediated rewards on intrinsic motivation," *Journal of Personality and Social Psychology*, 18(1): 105–115.

Deci, E. L. and R. M. Ryan (1971). "Effects of externally mediated rewards on intrinsic motivation," *Journal of Personality and Social Psychology*, 18: 105–115.

Delin, Catherine R and Roy F Baumeister (1994). "Praise: More than just social reinforcement," *Journal for the Theory of Social Behaviour*, 24(3): 219–241.

DellaVigna, Stefano and Devin Pope (2017). "What motivates effort? Evidence and expert forecasts," *The Review of Economic Studies*: rdx033.

Dur, Robert, Arjan Non, and Hein Roelfsema (2010). "Reciprocity and incentive pay in the workplace," *Journal of Economic Psychology*, 31(4): 676–686.

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

Farson, Richard E (1963). "Praise reappraised." *Harvard Business Review*.

Gneezy, Uri and Aldo Rustichini (2000). "Pay enough or don't pay at all," *Quarterly Journal of Economics*, 115(3): 791–810.

Goerg, Sebastian J and Sebastian Kube (2012). "Goals (th) at work," *Preprints of the Max Planck Institute for Research on Collective Goods*, 19.

Grinstein, Amir and Ann Kronrod (2016). "Does sparing the rod spoil the child? How praising, scolding, and an assertive tone can encourage desired behaviors," *Journal of Marketing Research*, 53(3): 433–441.

Henderlong, Jennifer and Mark R Lepper (2002). "The effects of praise on children's intrinsic motivation: A review and synthesis," *Psychological Bulletin*, 128(5): 774–795.

Hermalin, Benjamin E (2015). "At the helm, Kirk or Spock? The pros and cons of charismatic leadership," Working paper.

Holmström, Bengt and Paul Milgrom (1991). "Multitask principal-agent analyses: Incentive contracts, asset ownership, and job design," *Journal of Law, Economics, and Organization*, 7(Special Issue): 24–52.

Hong, Fuhai, Tanjim Hossain, John A List, and Migiwa Tanaka (2013). "Testing the theory of multitasking: Evidence from a natural field experiment in Chinese factories," *NBER working paper 19660*.

Horton, John J (2010). "Online labor markets," *Internet and Network Economics*: 515–522.

Horton, John J., David G. Rand, and Richard J. Zeckhauser (2011). "The online laboratory: Conducting experiments in a real labor market," *Experimental Economics*, 14: 399–425.

Judge, Timothy A and Joyce E Bono (2000). "Five-factor model of personality and transformational leadership," *Journal of Applied Psychology*, 85(5): 751.

Kässi, Otto and Vili Lehdonvirta (2016). "Online labour index: Measuring the online gig economy for policy and research," Working Paper 74943, Munich Personal RePEc Archive.

Kvaløy, Ola, Petra Nieken, and Anja Schöttner (2015). "Hidden benefits of reward: A field experiment on motivation and monetary incentives," *European Economic Review*, 76: 188–199.

Kvaløy, Ola and Anja Schöttner (2015). "Incentives to motivate," *Journal of Economic Behavior & Organization*, 116: 26–42.

Levenshtein, V. I. (1966). "Binary Codes Capable of Correcting Deletions, Insertions and Reversals," *Soviet Physics Doklady*, 10: 707.

List, John A. and Fatemeh Momeni (2017). "When corporate social responsibility backfires: Theory and evidence from a natural field experiment," Working Paper 24169, National Bureau of Economic Research.

Locke, E. A. and G. P. Latham (1984). *Goal Setting: A Motivational Technique That Works*, Englewood Cliffs, NJ:Prentice-Hall.

Locke, E. A. and G. P. Latham (2002). "Building a practically useful theory of goal setting and task motivation: A 35-year odyssey," *American Psychologist*, 57: 705–717.

Mason, Winter and Siddharth Suri (2012). "Conducting behavioral research on Amazon's Mechanical Turk," *Behavior Research Methods*, 44(1): 1–23.

Paolacci, Gabriele and Jesse Chandler (2014). "Inside the turk: Understanding Mechanical Turk as a participant pool," *Current Directions in Psychological Science*, 23(3): 184–188.

Paolacci, Gabriele, Jesse Chandler, and Panagiotis G. Ipeirotis (2010). "Running experiments on Amazon Mechanical Turk," *Judgment and Decision making*, 5(5): 411–419.

Peer, Eyal, Laura Brandimarte, Sonam Samat, and Alessandro Acquisti (2017). "Beyond the Turk: Alternative platforms for crowdsourcing behavioral research," *Journal of Experimental Social Psychology*, 70: 153–163.

Pokorny, K. (2008). "Pay – but don't pay too much: An experimental study on the impact of incentives," *Journal of Economic Behavior and Organization*, 66(2): 251–264.

Purvanova, Radostina K. and Joyce E. Bono (2009). "Transformational leadership in context: Face-to-face and virtual team," *The Leadership Quarterly*, 20: 343–357.

de Quidt, Jonathan (2017). "Your loss is my gain: A recruitment experiment with framed incentives," *Journal of the European Economic Association*: jvx016.

Robbins, S. P. and T. A. Judge (2013). *Organizational Behavior*, Pearson Education Limited, Harlow.

Rotemberg, Julio J and Garth Saloner (2000). "Visionaries, managers, and strategic direction," *RAND Journal of Economics*: 693–716.

Shamir, Boas, Robert J House, and Michael B Arthur (1993). "The motivational effects of charismatic leadership: A self-concept based theory," *Organization Science*, 4(4): 577–594.

Van den Steen, Eric (2005). "Organizational beliefs and managerial vision," *Journal of Law, Economics, and Organization*, 21(1): 256–283.

Tims, Maria, Arnold B Bakker, and Despoina Xanthopoulou (2011). "Do transformational leaders enhance their followers' daily work engagement?" *The Leadership Quarterly*, 22(1): 121–131.

Walumbwa, Fred O, Bruce J Avolio, and Weichun Zhu (2008). "How transformational leadership weaves its influence on individual job performance: The role of identification and efficacy beliefs," *Personnel Psychology*, 61(4): 793–825.

Wang, G., I. Oh, S. H. Courtright, and A. E. Colbert (2011). "Transformational leadership and performance across criteria and levels: A meta-analytic review of 25 years of research," *Group and Organization Management*, 36: 223–270.

Table 1: Treatment table

| | Leadership technique | | | |
|---|---|---|---|---|
| Performance pay | Neutral | Expectations | Goal | All |
| No piece rate | 300 | 292 | 299 | 891 |
| Low piece rate | 295 | 301 | 295 | 891 |
| High piece rate | 302 | 297 | 299 | 898 |
| All | 897 | 890 | 893 | 2680 |

*Note*: The table gives an overview of the experimental design and shows the combination of the monetary and non-monetary treatment interventions. The number of subjects for each treatment cell is indicated as well.

Table 2: Background characteristics of subjects

| | | Age | Female | Education | Mobile device | Latin | |
|---|---|---|---|---|---|---|---|
| **Performance pay** | **Leadership technique** | **Mean (se)** | **Mean (se)** | **Mean (se)** | **Mean (se)** | **Mean (se)** | **N** |
| No piece rate | Neutral | 36.28 (0.59) | 0.50 (0.03) | 3.12 (0.08) | 0.05 (0.01) | 1.42 (0.04) | 300 |
| | Expectations | 36.04 (0.62) | 0.50 (0.03) | 3.24 (0.08) | 0.03 (0.01) | 1.38 (0.04) | 292 |
| | Goal | 35.77 (0.65) | 0.54 (0.03) | 3.12 (0.07) | 0.07 (0.01) | 1.44 (0.04) | 299 |
| Low piece rate | Neutral | 35.87 (0.64) | 0.50 (0.03) | 3.08 (0.07) | 0.07 (0.01) | 1.41 (0.04) | 295 |
| | Expectations | 34.49 (0.56) | 0.50 (0.03) | 3.07 (0.08) | 0.04 (0.01) | 1.41 (0.04) | 301 |
| | Goal | 35.42 (0.64) | 0.49 (0.03) | 3.15 (0.08) | 0.03 (0.01) | 1.45 (0.05) | 295 |
| High piece rate | Neutral | 34.93 (0.61) | 0.46 (0.03) | 3.02 (0.08) | 0.05 (0.01) | 1.46 (0.04) | 302 |
| | Expectations | 35.15 (0.64) | 0.52 (0.03) | 3.13 (0.07) | 0.06 (0.01) | 1.40 (0.04) | 297 |
| | Goal | 36.08 (0.65) | 0.54 (0.03) | 3.09 (0.08) | 0.05 (0.01) | 1.47 (0.05) | 299 |
| **All** | | 35.56 (0.21) | 0.50 (0.01) | 3.11 (0.03) | 0.05 (0.00) | 1.43 (0.01) | 2680 |

*Note*: The table reports background characteristics of subjects participating in the experiment. Subjects were recruited through the Amazon Mechanical Turk crowd-sourcing platform. "Age" is a continuous variable measuring participants' age in years; "Female" captures the proportion of females; "Education" is an ordinal scaled variable: 1 = High School', 2 = Some College, 3 = 2 year College Degree, 4 = 4 year College Degree, 5 = Masters Degree, 6 = Doctoral Degree; "Mobile device" captures the share of mobile users; "Latin" is an ordinal scaled variable measuring the subject's knowledge of Latin: 1 = Not at all, 5 = Very well.

Table 3: Treatment effects on quantity

| Column | I | II | III | IV |
|---|---|---|---|---|
| Low piece rate | 1.387*** | | 1.950** | 1.990*** |
| | (0.423) | | (0.780) | (0.750) |
| High piece rate | 1.415*** | | 2.190*** | 1.983*** |
| | (0.415) | | (0.737) | (0.719) |
| Expectations | | -1.206*** | -1.075 | -1.280* |
| | | (0.440) | (0.730) | (0.704) |
| Goal | | -0.333 | 0.847 | 0.794 |
| | | (0.417) | (0.690) | (0.670) |
| Low piece rate | | | -0.272 | -0.474 |
| × Expectations | | | (1.082) | (1.041) |
| High piece rate | | | -0.153 | 0.099 |
| × Expectations | | | (1.045) | (1.017) |
| Low piece rate | | | -1.379 | -1.653* |
| × Goal | | | (1.020) | (0.976) |
| High piece rate | | | -2.160** | -1.987** |
| × Goal | | | (0.999) | (0.968) |
| Age | | | | -0.192*** |
| | | | | (0.015) |
| Female | | | | 0.706** |
| | | | | (0.334) |
| Education | | | | 0.549*** |
| | | | | (0.130) |
| Mobile device | | | | -5.019*** |
| | | | | (0.866) |
| Latin | | | | 0.404 |
| | | | | (0.251) |
| Constant | 22.209*** | 23.656*** | 22.277*** | 26.871*** |
| | (0.290) | (0.316) | (0.505) | (0.870) |
| N | 2680 | 2680 | 2680 | 2680 |
| $R^2$ | 0.005 | 0.003 | 0.011 | 0.083 |

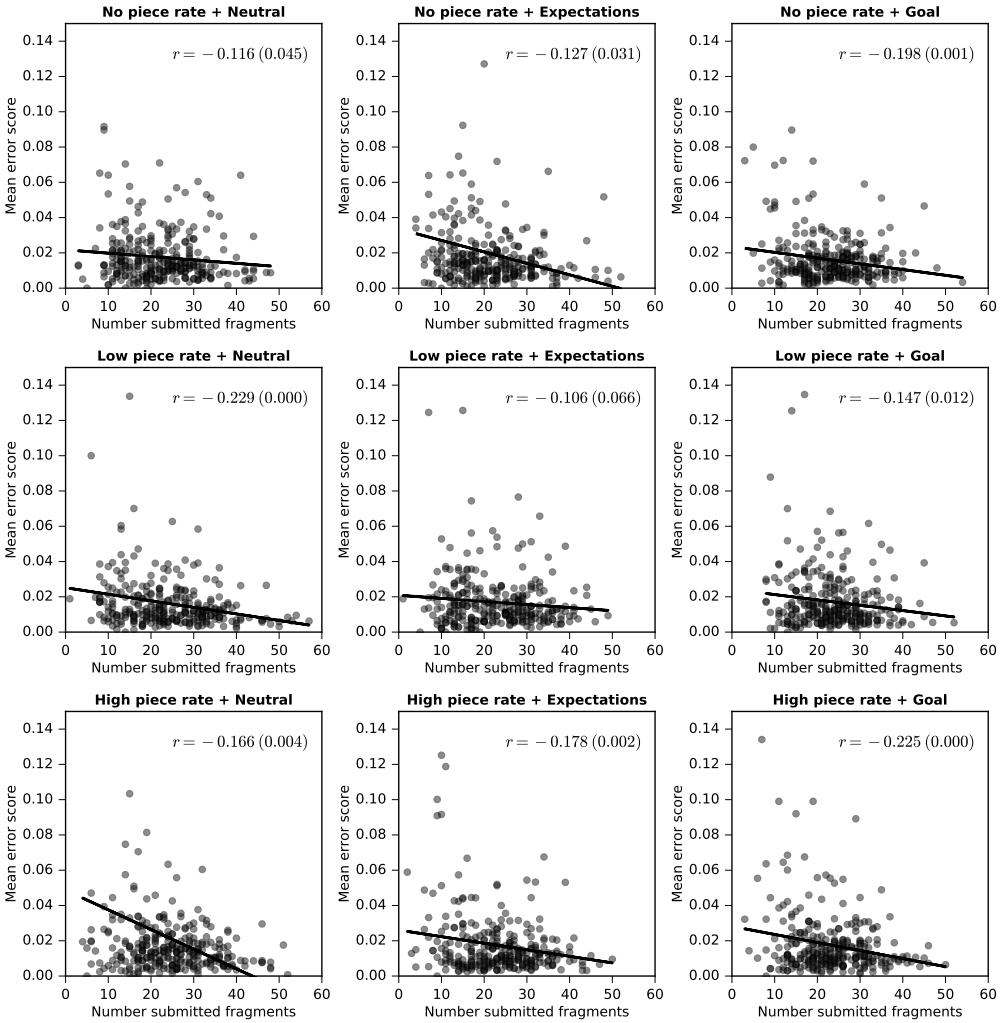
*Note*: The table reports linear regression results of the number of fragments submitted per worker on a set of explanatory variables. "Low piece rate": indicator variable taking the value one for the *Low piece rate* treatment. "High piece rate": indicator variable taking the value one for the *High piece rate* treatment. "Expectations": indicator variable taking the value one for the *Expectation* treatment. "Goal": indicator variable taking the value one for the *Goal* treatment. "Age": continuous variable measuring a worker's age. "Education" is an ordinal scaled variable: 1 = High School, 2 = Some College, 3 = 2 year College Degree, 4 = 4 year College Degree, 5 = Masters Degree, 6 = Doctoral Degree; "Female": indicator variable taking the value one if the worker is a female. "Mobile device": indicator variable taking the value one if the worker is a female. Robust standard errors in parentheses ($* : p < 0.1$, $** : p < 0.05$, $*** : p < 0.01$).

Table 4: Treatment effects on quality

| Column | I | II | III | IV |
|---|---|---|---|---|
| Low piece rate | -0.001 | | -0.001 | -0.002 |
| | (0.001) | | (0.002) | (0.002) |
| High piece rate | 0.001 | | 0.004 | 0.003 |
| | (0.002) | | (0.002) | (0.002) |
| Expectations | | -0.000 | 0.002 | 0.002 |
| | | (0.002) | (0.003) | (0.003) |
| Goal | | -0.001 | -0.002 | -0.001 |
| | | (0.001) | (0.002) | (0.002) |
| Low piece rate | | | -0.001 | -0.001 |
| × Expectations | | | (0.003) | (0.003) |
| High piece rate | | | -0.006 | -0.006 |
| × Expectations | | | (0.004) | (0.004) |
| Low piece rate | | | 0.003 | 0.003 |
| × Goal | | | (0.003) | (0.003) |
| High piece rate | | | -0.002 | -0.002 |
| × Goal | | | (0.003) | (0.003) |
| Age | | | | -0.000 |
| | | | | (0.000) |
| Female | | | | -0.004*** |
| | | | | (0.001) |
| Education | | | | -0.001* |
| | | | | (0.000) |
| Mobile device | | | | 0.008** |
| | | | | (0.003) |
| Latin | | | | 0.000 |
| | | | | (0.001) |
| Constant | 0.018*** | 0.018*** | 0.017*** | 0.021*** |
| | (0.001) | (0.001) | (0.002) | (0.003) |
| N | 62026 | 62026 | 62026 | 62026 |
| $R^2$ | 0.002 | 0.002 | 0.002 | 0.002 |
| $R^2$ (Within) | 0.000 | 0.000 | 0.000 | 0.000 |
| $R^2$ (Between) | 0.001 | 0.000 | 0.003 | 0.013 |

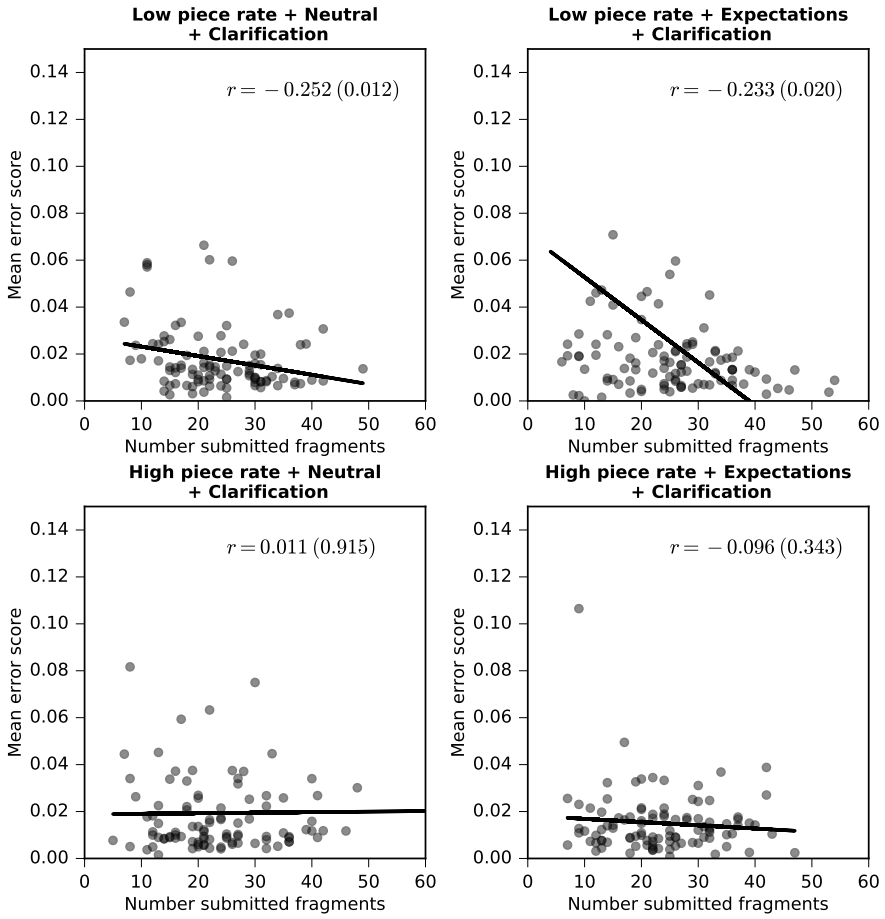
*Note*: The table reports random effects panel regression results of error score per fragment by a single worker on a set of explanatory variables. "Low piece rate": indicator variable taking the value one for the *Low piece rate* treatment. "High piece rate": indicator variable taking the value one for the *High piece rate* treatment. "Expectations": indicator variable taking the value one for the *Expectation* treatment. "Goal": indicator variable taking the value one for the *Goal* treatment. "Age": continuous variable measuring a worker's age. "Education" is an ordinal scaled variable: 1 = High School, 2 = Some College, 3 = 2 year College Degree, 4 = 4 year College Degree, 5 = Masters Degree, 6 = Doctoral Degree; "Female": indicator variable taking the value one if the worker is a female. "Mobile device": indicator variable taking the value one if the worker is a female. Standard errors in parentheses ($* : p < 0.1$, $** : p < 0.05$, $*** : p < 0.01$).

Figure 1: Quantity vs. Quality




*Note:* The figure plots the number of submitted fragments per worker against the timed average error score for all submitted fragments per worker for each treatment. Indicated as well are the overlaid linear predictions as well as the Pearson correlation coefficient along with p-values (in parentheses).

Figure 2: Quantity vs. Quality, clarification treatments only

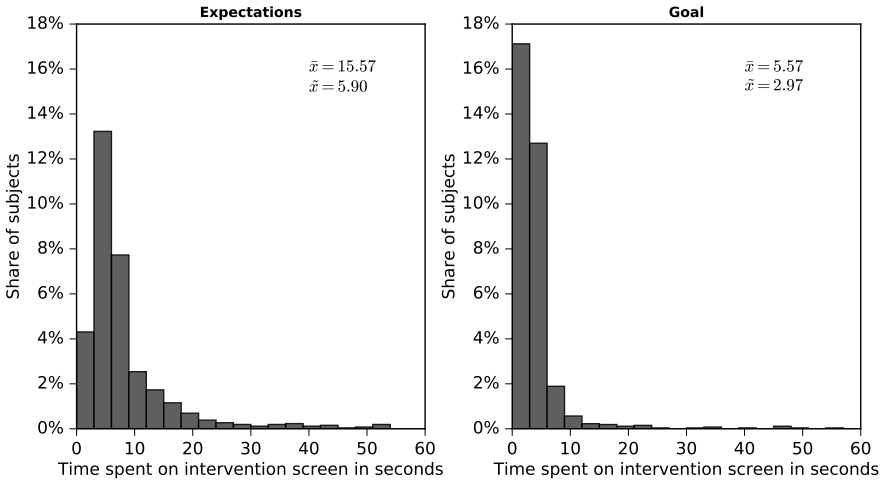



*Note:* The figure plots the number of submitted fragments per worker against the timed average error score for all submitted fragments per worker for each clarification treatment. Indicated as well are the overlaid linear predictions as well as the Pearson correlation coefficient along with p-values (in parentheses).

# Appendix

## 5.1 Additional tables and figures

Figure S1: Time spent on intervention screen




*Note:* The figure shows the histogram of time spent on the intervention screen for the *Expectations* (left panel) and *Goal* treatment (right panel). The mean ($\bar{x}$) and median ($\tilde{x}$) time spend on the intervention screen are reported in each panel as well.

Table S1: Quality-quantity trade-off

| Column | I | II | III | IV |
|---|---|---|---|---|
| No. Fragments | -0.0004*** | -0.0004*** | -0.0004*** | -0.0002* |
| | (0.0001) | (0.0001) | (0.0001) | (0.0001) |
| No piece rate+Expectations | | 0.0018 | | 0.0118 |
| | | (0.0028) | | (0.0107) |
| No piece rate+Goal | | -0.0011 | | 0.0018 |
| | | (0.0011) | | (0.0044) |
| Low piece rate+Neutral | | -0.0004 | | 0.0034 |
| | | (0.0013) | | (0.0041) |
| Low piece rate+Expectations | | -0.0003 | | -0.0009 |
| | | (0.0012) | | (0.0040) |
| Low piece rate+Goal | | 0.0003 | | 0.0026 |
| | | (0.0013) | | (0.0046) |
| High piece rate+Neutral | | 0.0049 | | 0.0269 |
| | | (0.0038) | | (0.0181) |
| High piece rate+Expectations | | 0.0004 | | 0.0043 |
| | | (0.0014) | | (0.0050) |
| High piece rate+Goal | | 0.0005 | | 0.0064 |
| | | (0.0013) | | (0.0047) |
| No piece rate+Expectations | | | 0.0000 | -0.0005 |
| × No. Fragments | | | (0.0001) | (0.0004) |
| No piece rate+Goal | | | -0.0001 | -0.0001 |
| × No. Fragments | | | (0.0000) | (0.0002) |
| Low piece rate+Neutral | | | -0.0000 | -0.0002 |
| × No. Fragments | | | (0.0000) | (0.0001) |
| Low piece rate+Expectations | | | -0.0000 | 0.0000 |
| × No. Fragments | | | (0.0000) | (0.0002) |
| Low piece rate+Goal | | | -0.0000 | -0.0001 |
| × No. Fragments | | | (0.0000) | (0.0002) |
| High piece rate+Neutral | | | 0.0001 | -0.0009 |
| × No. Fragments | | | (0.0001) | (0.0006) |
| High piece rate+Expectations | | | -0.0000 | -0.0002 |
| × No. Fragments | | | (0.0001) | (0.0002) |
| High piece rate+Goal | | | -0.0000 | -0.0003 |
| × No. Fragments | | | (0.0000) | (0.0002) |
| Constant | 0.0281*** | 0.0275*** | 0.0281*** | 0.0217*** |
| | (0.0025) | (0.0022) | (0.0026) | (0.0028) |
| N | 2680 | 2680 | 2680 | 2680 |
| $R^2$ | 0.018 | 0.021 | 0.019 | 0.029 |

*Note*: The table reports linear regression results of the averaged error score for all fragments on the number of fragments submitted per worker ("No. Fragments"). The remaining variables are indicator variables for treatments. Variables with an "×" operator represent interaction variables for treatments in combination with the number of fragments submitted per worker. Robust standard error in parentheses ($* : p < 0.1$, $** : p < 0.05$, $*** : p < 0.01$).

Table S2: Quality-quantity trade-off, clarification treatments only

| Column | I | II | III | IV |
|---|---|---|---|---|
| No. Fragments | -0.0006 | -0.0006 | -0.0007 | -0.0004** |
| | (0.0004) | (0.0005) | (0.0005) | (0.0002) |
| Low piece rate+Expectations | | 0.0088 | | 0.0438 |
| | | (0.0088) | | (0.0447) |
| High piece rate+Neutral | | 0.0018 | | -0.0083 |
| | | (0.0027) | | (0.0062) |
| High piece rate+Expectations | | -0.0027 | | -0.0089 |
| | | (0.0020) | | (0.0069) |
| Low piece rate+Expectations | | | 0.0001 | -0.0014 |
| × No. Fragments | | | (0.0001) | (0.0015) |
| High piece rate+Neutral | | | 0.0001 | 0.0004* |
| × No. Fragments | | | (0.0001) | (0.0002) |
| High piece rate+Expectations | | | -0.0001 | 0.0003 |
| × No. Fragments | | | (0.0001) | (0.0002) |
| Constant | 0.0344*** | 0.0329*** | 0.0348*** | 0.0271*** |
| | (0.0125) | (0.0108) | (0.0127) | (0.0044) |
| N | 396 | 396 | 396 | 396 |
| $R^2$ | 0.019 | 0.028 | 0.021 | 0.058 |

*Note*: The table reports linear regression results of the averaged error score for all fragments on the number of fragments submitted per worker ("No. Fragments"). The remaining variables are indicator variables for treatments. Variables with an "×" operator represent interaction variables for treatments in combination with the number of fragments submitted per worker. Robust standard error in parentheses ($* : p < 0.1$, $** : p < 0.05$, $*** : p < 0.01$).

Table S3: Quality-quantity trade-off, clarification vs. no clarification

| Column | I | II | III |
|---|---|---|---|
| No. Fragments | -0.0005*** | -0.0005*** | -0.0005*** |
| | (0.0002) | (0.0001) | (0.0001) |
| Clarification | | 0.0047 | 0.0041 |
| | | (0.0132) | (0.0132) |
| No. Fragments × Clarification | | -0.0001 | -0.0001 |
| | | (0.0005) | (0.0005) |
| Constant | 0.0309*** | 0.0297*** | 0.0384*** |
| | (0.0046) | (0.0044) | (0.0055) |
| Controls | No | No | Yes |
| N | 1591 | 1591 | 1591 |
| R2 | 0.018 | 0.019 | 0.026 |

*Note*: The table reports linear regression results of the averaged error score for all fragments on the number of fragments submitted per worker ("No. Fragments"). "Clarification": indicator variable taking the value one for the *Clarification* treatments. "No. Fragments × Clarification" is an interaction indicator variable for the number of fragments submitted per worker in the *Clarification* treatments. Controls include variables for workers' age, gender, education, use of mobile device and knowledge of Latin. Robust standard error in parentheses ($* : p < 0.1$, $** : p < 0.05$, $*** : p < 0.01$).

Table S4: Treatment effects on quantity, with clarification treatments

| Column | I | II | III | IV | V |
|---|---|---|---|---|---|
| High piece rate | 0.094 | | | 0.240 | 0.005 |
| | (0.479) | | | (0.801) | (0.770) |
| Expectations | | -0.863* | | -1.347* | -1.758** |
| | | (0.479) | | (0.799) | (0.767) |
| Clarification | | | 0.269 | -0.543 | -0.712 |
| | | | (0.564) | (1.068) | (1.032) |
| High piece rate × Expectations | | | | 0.119 | 0.550 |
| | | | | (1.095) | (1.062) |
| High piece rate × Clarification | | | | -0.087 | 0.469 |
| | | | | (1.600) | (1.528) |
| Expectations × Quality concern | | | | 2.473 | 2.732* |
| | | | | (1.601) | (1.557) |
| High piece rate × Expectations × Clarification | | | | -1.532 | -2.242 |
| | | | | (2.257) | (2.180) |
| Constant | 23.723*** | 24.203*** | 23.703*** | 24.227*** | 30.083*** |
| | (0.346) | (0.346) | (0.274) | (0.594) | (1.126) |
| Controls | No | No | No | No | Yes |
| N | 1591 | 1591 | 1591 | 1591 | 1591 |
| $R^2$ | 0.000 | 0.002 | 0.000 | 0.004 | 0.073 |

*Note*: The table reports linear regression results of the number of fragments submitted per worker on a set of explanatory variables. "High piece rate": indicator variable taking the value one for the *High piece rate* treatment. "Expectations": indicator variable taking the value one for the *Expectation* treatment. "Clarification": indicator variable taking the value one for the *Clarification* treatments. Controls include variables for workers' age, gender, education, use of mobile device and knowledge of Latin. Robust standard errors in parentheses ($* : p < 0.1$, $** : p < 0.05$, $*** : p < 0.01$).

Table S5: Treatment effects on quality, with clarification treatments

| Column | I | II | III | IV | V |
|---|---|---|---|---|---|
| High piece rate | 0.001 | | | 0.005 | 0.005 |
| | (0.002) | | | (0.003) | (0.003) |
| Expectations | | -0.001 | | 0.001 | 0.001 |
| | | (0.002) | | (0.003) | (0.003) |
| Clarification | | | 0.001 | 0.002 | 0.002 |
| | | | (0.002) | (0.004) | (0.004) |
| High piece rate × Expectations | | | | -0.005 | -0.005 |
| | | | | (0.005) | (0.005) |
| High piece rate × Clarification | | | | -0.003 | -0.004 |
| | | | | (0.005) | (0.005) |
| Expectations × Quality concern | | | | 0.007 | 0.006 |
| | | | | (0.007) | (0.007) |
| High piece rate × Expectations × Clarification | | | | -0.007 | -0.007 |
| | | | | (0.009) | (0.009) |
| Constant | 0.018*** | 0.019*** | 0.018*** | 0.016*** | 0.021*** |
| | (0.001) | (0.001) | (0.001) | (0.002) | (0.004) |
| Controls | No | No | No | No | Yes |
| N | 37818 | 37818 | 37818 | 37818 | 37818 |
| $R^2$ | 0.002 | 0.002 | 0.002 | 0.002 | 0.002 |
| $R^2$ (Within) | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| $R^2$ (Between) | 0.000 | 0.000 | 0.001 | 0.005 | 0.013 |

*Note*: The table reports random effects panel regression results of error score per fragment by a single worker on a set of explanatory variables. "High piece rate": indicator variable taking the value one for the *High piece rate* treatment. "Expectations": indicator variable taking the value one for the *Expectation* treatment. "Clarification": indicator variable taking the value one for the *Clarification* treatments. Controls include variables for workers' age, gender, education, use of mobile device and knowledge of Latin. Standard errors in parentheses ($* : p < 0.1$, $** : p < 0.05$, $*** : p < 0.01$).

## 5.2 Instructions

You will be paid a **fixed compensation of $2** for working on this project. [piece rate Treatments: In addition, you will receive **a bonus of $0.01 ($0.05)** for each completed fragment.] The compensation will be sent to you within two days after the completion of this HIT.

[Approval treatments : Once you have completed the HIT, you will be approved automatically, which means that your performance will **not affect your approval rate**.][15]

[Clarification Treatments: In order to pay the bonus in due time, we pay it for submitted fragments without controlling for typing errors. Once you have completed the HIT, you will be approved automatically, which means that your performance will **not affect your approval rate**.][16]

{New page}

Please read the instructions below carefully. In the assignment you will be shown fragments of an ancient Latin text. You are asked to type the text into the blank space below the fragment using your keyboard. If you can't read a specific letter, please insert a question mark instead of the letter.

Below, you see an example of the task. In the actual assignment, after you have submitted the text, a new fragment will appear on your screen. In total, you will have to work on the assignment for 10 minutes. We ask you to complete as many fragments as possible.

After finishing the assignment, you will be taken to a short questionnaire.

[Example fragment here]

{New page for Expectation treatments}

Before you start, we want to emphasize how happy we are that you've decided to work for us. You've proven to be a successful and diligent worker on MTurk with an impressive approval rate!

[15] These treatments where pooled with the main treatments, compare fn. 4.
[16] We explain these treatments in Section 4.

{New page for Goal treatments}

Efficient work is important. Please try to submit at least 25 fragments.