ECONSTOR
Make Your Publications Visible.

A Service of

ZBW Leibniz-Informationszentrum Wirtschaft
Leibniz Information Centre for Economics

Fest, Sebastian; Kvaloy, Ola; Nieken, Petra; Schöttner, Anja

Conference Paper

Motivation and Incentives in an Online Labor Market

Beiträge zur Jahrestagung des Vereins für Socialpolitik 2020: Gender Economics

Provided in Cooperation with:
Verein für Socialpolitik / German Economic Association

Suggested Citation: Fest, Sebastian; Kvaloy, Ola; Nieken, Petra; Schöttner, Anja (2020) : Motivation and Incentives in an Online Labor Market, Beiträge zur Jahrestagung des Vereins für Socialpolitik 2020: Gender Economics, ZBW - Leibniz Information Centre for Economics, Kiel, Hamburg

This Version is available at:
https://hdl.handle.net/10419/224586

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.

WWW.ECONSTOR.EU

Mitglied der Leibniz-Gemeinschaft

Motivation and Incentives in an Online Labor Market

Sebastian Fest Ola Kvaløy Petra Nieken
Anja Schöttner*

Abstract

An increasing number of workers participate in online labor markets. In contrast to traditional employment relationships within firms, the interaction between online workers and their employers are short and impersonal, which makes motivating online workers more challenging. We present results from two large-scale real-effort experiments on Amazon Mechanical Turk investigating the effects of monetary and non-monetary motivational instruments. In the first experiment, we study the effects of performance pay and simple upfront messages (praise or reference points) on performance. The second experiment concentrates on the effects of communication techniques used by charismatic leaders. Performance pay increases output significantly. Sending simple messages, however, can have a significantly negative effect on output. The results from the second experiment show that charismatic communication techniques can also backfire when only a subset of them is used, whereas using a broad set including quantitative goals increases output significantly. Neither intervention had any effect on the quality of work.

Keywords: Online Labor Market, Performance Pay, Motivation, Charismatic Leadership

*The authors are listed in alphabetical order. All authors contributed equally to the project. Fest: Norwegian School of Economics, FAIR, 5045 Bergen, Norway (e-mail: sebastian.fest@nhh.no); Kvaløy: University of Stavanger, UiS Business School, 4036 Stavanger, Norway (e-mail: ola.kvaloy@uis.no); Nieken: Karlsruhe Institute of Technology, Institute of Management, Kaiserstr. 89, 76133 Karlsruhe, Germany (e-mail: petra.nieken@kit.edu); Schöttner: Humboldt-Universität zu Berlin, School of Business and Economics, 10178 Berlin, Germany (e-mail: anja.schoettner@hu-berlin.de). 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 and the Deutsche Forschungsgemeinschaft through CRC TRR 190 (project number 280092119) is gratefully acknowledged.

1 Introduction

Online labor markets give rise to a new form of labor. Millions of online workers sell everything from complex consulting services to simple production and routine jobs through platforms such as Elance-oDesk, Eden McCallum, or Amazon Mechanical Turk. Over one-third of U.S. workers participate in the so-called gig economy, either through their primary or secondary jobs (Gallup, 2018). The world-wide annual growth of the so-called 'gig economy' has been estimated to be 14% (Kässi and Lehdonvirta, 2018). The International Labour Organization (ILO) regards the emergence of online digital labor platforms as the major transformation of work life in the last decade (ILO, 2018). Advancements in information and communication technologies have dramatically lowered the transaction costs of using online markets, and this trend can be expected to continue, suggesting that both firms and workers will use online labor markets even more in the future (Coase, 1937; Munger, 2015). Online labor platforms have not only disrupted existing business models, but also fundamentally changed employment relationships. In contrast to employees within firms, online workers typically have spot contracts with many different employers. They usually work from home, and do not have any personal contact with employers or colleagues.

In traditional employment relationships, workers are typically motivated by a well-balanced combination of transactional and transformational leadership techniques (Zehnder, Herz, and Bonardi, 2017). Transactional leadership corresponds to the exchange of resources through the design of contracts, incentive systems, and organizational institutions. Transformational leaders provide a vision, define meaning and goals, praise performance, or employ rhetorical techniques. Transactional leadership techniques are straightforward to implement in online labor markets, e.g., by granting performance related pay. By contrast, providing effective transformational leadership for online workers seems to be a much bigger challenge.

In firms, transformational leadership techniques can be employed in regular face-to-face interactions. In online labor markets, employers are typically left with short digital messages that lack non-verbal elements such as visual or auditory cues, which are main carriers of emotional communication (Purvanova and Bono, 2009). Antonakis, Fenley, and Liechti (2011) show that charismatic leaders usually use both, verbal and non-verbals clues, to reach out to their followers, which calls the power of simple written messages into question. More generally, communication theories stress the superiority of face-to-face communication over computer-mediated communication (see Purvanova and Bono (2009) for an overview).

Even though transformational leadership faces severe challenges in online labor markets, it can be particularly valuable in this domain. Tasks performed in online labor markets are often simple, and it is sometimes argued that transformational leadership is primarily needed in complex work environments (Zehnder et al., 2017). However, also performance in simple tasks suffers from problems that arise in purely transactional employment arrangements due to the incom-

pleteness of contracts. Also simple tasks have different dimensions of whom not all are easily measurable, e.g., quantity versus quality of work. Performance pay that rewards workers for the easily measurable dimensions may distract workers' attention from the less easily measurable, but also important, dimensions (Holmström and Milgrom, 1991).

Employing transformational leadership techniques may be more effective at motivating workers and less costly than granting performance pay. Also, execution of simple tasks is often boring and workers might not directly see the inherent value of their work. In particular for such tasks it may be essential to provide workers with vision, meaning, and a clear purpose of work. Relatedly, e-leadership research asserts that transformational leadership can be particularly effective in the virtual domain, e.g., due to the need for leadership in ambiguous environments (e.g., Avolio, Kahai, and Dodge, 2000; Purvanova and Bono, 2009). The e-leadership literature, however, focuses on how to influence members of virtual project teams who regularly and repeatedly engage in computer-mediated communication with one another and their leader, and thus considers a setting that crucially differs from one-shot interactions between employers and workers in online labor markets.

Up to now, we lack a systematic understanding whether and how transformational motivational techniques and in particular communication techniques can be successfully used in online labor markets. We present two empirical studies that contribute to filling this gap by investigating performance effects of written communication in an online labor market. To conceptualize transformational leadership, we apply the economic theory-based definition by Zehnder et al. (2017), who propose that transformational leaders affect workers' behavior by shaping their beliefs or preferences. Leaders induce others to choose desired actions without enforcing rules or exercising coercion but inspire through their words and behavior (Hermalin, 1998). Our study sheds light on whether employers in online labor markets can affect performance through such "soft" leadership techniques. More specifically, we are interested in the effects of (i) simple upfront messages that may exploit workers' reciprocal or reference-dependent preferences and (ii) written communication that utilizes charismatic leadership tactics which are known to be successful in face-to-face interactions (Antonakis et al., 2011; Antonakis, d'Adda, Weber, and Zehnder, 2019; Meslec, Curseu, Fodor, and Kenda, 2020).

Antonakis, Bastardoz, Jacquart, and Shamir (2016) define charisma as "values-based, symbolic, and emotion-laden leader signaling." Charismatic and transformational leadership can be seen as two distinct concepts that partly overlap. By employing charismatic leadership tactics, a leader could but does not have to have a transformational effect on followers. In particular, charismatic communication may be able to "transform" people's beliefs or preferences as described by Zehnder et al. (2017). We follow the definition provided by Antonakis et al. (2016) because it offers a solid basis for testing and reliably coding aspects of leadership communication in our setting.

In our first study, we address the question if simple upfront written messages,

expressing praise or communicating reference points, affect performance. Moreover, we ask whether and how contingent monetary rewards affect performance or interact with written messages. For employers, who seek for the optimal combination of all available motivational devices, it is important to understand whether non-monetary motivational techniques make monetary rewards more or less effective. In our second study, we focus on charismatic communication tactics and investigate whether these tactics also work in online labor markets when only presented in writing. We also disentangle the effects of quantitative goals and goal-related tactics and other charismatic communication tactics because goals are often used in isolation practice.

Our results help to extend leadership research to a setting where there is no organizational context and no repeated interaction between employer and worker. We ask the question if non-monetary interventions and leadership communication tools can be effective in such impersonal settings. If the answer is positive, using these techniques can lead to better work results and reduce the costs of labor for the employer relative to performance-based pay. In addition, we can investigate if there is an interaction of transactional techniques in the form of monetary rewards and softer and cheaper non-monetary interventions. Our study clarifies if charismatic leadership can be effective in the absence of non-verbal signals such as body language and tone of voice. Our results also enhance our understanding of the interaction between goals and other forms of verbal charismatic leadership communication. In particular, we explore if a broad set of verbal charismatic leadership tactics is needed to raise performance in an online setting. Due to the nature of our task, our results offer insights into both, quantitative effects and qualitative effects of our interventions.

Our results primarily inform the large and steadily growing number of online employers who seek productive workers. It may, however, to some extent also inform virtual leadership outside the online labor market domain. Nowadays, people work from home more frequently than in the past, and digital communication and online meetings substitute for physical presence also in traditional employment relationships. The Covid-19 pandemic has further boosted home office practice, and it is expected that companies will at least partially continue their new work practices after the pandemic (The Economist, 2020). Leaders are thus increasingly expected to motivate their workforce digitally, with fewer communication techniques at their disposal.

Our paper contributes to the empirical leadership literature in several aspects: We present a series of large-scale field experiments that allow for causal inference. We employ full randomization and did not deceive our workers. They worked on a real human intelligence task (HIT) on Amazon Mechanical Turk, for which we paid them. Leadership has been extensively studied in psychology and management, which has led to numerous important insights, but many studies exhibit methodological problems that confine feasible conclusions. In particular, field studies predicting outcomes from measured leadership styles typically do not use experimental interventions, but rather measure the effect of endogenously determined leadership styles (Antonakis, Bendahan, Jacquart, and Lalive,

2010). More specifically, there is not much evidence that varying how a leader communicates has a causal impact on workers' performance. Our paper thus contributes to the small literature that identifies casual effects of leader communication on worker behavior. Antonakis et al. (2019) and Meslec et al. (2020) show that charismatic leadership tactics can raise workers' output when workers listen to a leader's speech, which was delivered either in person or via video including non-verbal tactics. Kvaløy, Nieken, and Schöttner (2015) demonstrate that a simple motivational speech in a face-to-face setting increases both quantity and quality of output, but only if workers also receive performance pay. In contrast to these papers, we study the effects of monetary incentives and softer non-monetary techniques including goal setting and charismatic communication tactics in an online labor market where workers typically receive only written messages from the leader.

In simple work settings such as ours, specific and challenging (yet attainable) goals are considered to be effective motivational instruments in the fields of psychology and management (e.g., Locke and Latham, 1984, 2002) and economics (e.g., Goerg and Kube, 2012). Moreover, goal setting is a tactic employed by charismatic leaders (Antonakis et al., 2011). 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, Gómez-Miñambres, and Hernán-Gonzalez, 2015, 2018). Our paper sheds light on whether and how leaders in online labor markets can motivate workers by providing a reference point for their output.

Our paper further enhances our knowledge about the effectiveness of monetary incentives, in particular regarding their potential interaction with softer non-monetary techniques. Previous papers have found positive (Kvaløy et al., 2015) or no interactions (Kosfeld, Neckermann, and Yang, 2017; Meslec et al., 2020). On a more general level, our study is related to an increasing number of recent papers that study work incentives or participation decisions in online labor markets (e.g., Chandler and Kapelner, 2013; DellaVigna and Pope, 2018; de Quidt, 2018; Farrell, Grenier, and Leiby, 2017; List and Momeni, 2017).

The remainder of the paper is organized as follows. In the next section, we provide information on the labor market platform that we have used for our study. In Section 3 and 4, we present hypotheses, design and results of study 1 and study 2, respectively.[1] Section 5 provides a general discussion of the results. Finally, Section 6 concludes.

2 Online labor market platform

To study behavior in an online labor market, we chose to conduct our studies on Amazon Mechanical Turk (MTurk), one of the most prominent and widely used platforms that currently exist (Peer, Brandimarte, Samat, and Acquisti,

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

2017). MTurk offers firms the opportunity to outsource small, manual tasks to a large number of online workers. Potential employers, called "requesters," 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 (the so-called approval rate), 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. 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. An approval rate of 98% is often deemed critical in this regard among workers and employers.

There is typically no communication between worker and employer besides some basic information on the work task that employers post on MTurk and more specific instructions about the task that employers provide once the workers has accepted the task. If needed, workers can contact the employer via email and it is up to the employer to answer the requests or not.

Workers receive their payment through the platform from the employer. The employer freely decides about the amount of payment he or she is willing to offer. This payment will be announced in the job description posted on MTurk. If the employer accepts the work, the workers account will be credited with the respective payment. Employers can offer a fixed payment for a task and also assign bonus payments to workers to reward exceptional performance. In addition, employers are also able to assign a qualification to workers and offer future work only to workers with this qualification level. Other mechanisms for rewarding and motivating workers are not part of the platform.

3 Study 1

3.1 Aim and hypotheses study 1

In study 1, we investigate how work performance is affected by performance pay in the form of piece rates and non-monetary motivational techniques in the form of short upfront messages. We use a text transcription task, which is a typical task on MTurk that allows us to measure both quantity and quality of output. We can thus study potentially diverging effects of our interventions on the two different performance dimensions (see subsection 3.2 for a detailed description).

All workers obtain a fixed wage for participating in the study. In addition, they receive either no piece rate, a low piece rate, or a high piece rate. We use

two different piece rates to investigate the relationship between piece rates and performance, and in particular how this relationship depends on the height of the piece rate. Workers further receive either no upfront motivational message, an upfront message that praises workers' past performance based on their approval rates, or an upfront message that establishes a reference point regarding the quantity of output. Messages are displayed after the task description and before the work phase because this form of communication from employers to workers is most straightforward on the platform.

Our hypotheses are as follows: Paying workers more for achieving a higher output should motivate them to worker harder. Thus, keeping the upfront message constant, we expect output quantity to be increasing in the piece rate. However, if workers want to increase their payment under a piece rate, they will have to work faster, which may result in a lower output quality. Moreover, workers may intentionally shirk on quality to obtain higher payments. Using piece rates can thus lead to a multitasking problem (Holmström and Milgrom, 1991). Therefore, we hypothesize that, keeping the upfront message constant, output quality is decreasing in the piece rate. We further expect to find a negative correlation between quantity and quality when we pay a piece rate but not under a fixed wage.

By praising workers, the employer expresses recognition and appreciation for the workers. Workers may feel the need to reciprocate the friendly gesture by doing a good job (Falk and Fischbacher, 2006). Moreover, by referring to high approval rates, employers remind workers of their past good performance and workers might feel the need to live up to that implicit expectation of good work. We therefore hypothesize that, keeping the payment scheme constant, praise enhances performance in both dimensions, quality and quantity of delivered work.

When we provide a reference point, workers are asked to submit 25 fragments. We have chosen this output quantity based on data of the treatments without performance pay and no message intervention where workers managed to type 22.28 fragments on average. Achieving 25 fragments translates into belonging to the 39% of best performers. We hypothesize that providing this reference point increases the output quantity as workers want to reach the reference point of 25 fragments compared to treatments without a reference point, keeping the payment scheme constant. In economic terms, providing a reference point may trigger reference-dependent utility, which means that, everything else equal, workers experience a higher utility when they reach the reference point compared to when they produce less than the reference point (Corgnet et al., 2015, 2018). However, as argued above, working faster may lead to lower quality of work. Thus, quality should decrease when a reference point is provided.

We are also interested in the interaction effects between performance pay and upfront messages. Psychological theories of motivation predict that monetary incentives alone can crowd out intrinsic motivation and thereby weaken performance (e.g., Deci, 1971). However, recent behavioral economic 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, communication by a leader can help clarify the nature of the task or reveal more information about the personality and intentions of the leader. Indeed, Kvaløy et al. (2015) find in a field experiment that motivational talk enhances the effectiveness of performance pay. Following this line of argumentation, we hypothesize that upfront text messages and performance pay are complements also for the online workers we study.

3.2 Design study 1

3.2.1 Work task study 1

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 than others. The task therefore requires effort, attention, and diligence. An example fragment is given in Figure 1.

[Figure 1 about here]

3.2.2 Treatments study 1

We employed three different compensation schemes. Workers received either only a fixed wage of \$2, or, on top of the fixed wage, a low piece rate of \$0.01 or a high piece rate of \$0.05 per submitted fragment. 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."* A piece rate of 5 Cents leads to a considerable potential earnings increase compared to a piece rate of only 1 Cent. For example, a worker who submits 25 fragments yields a \$1 higher payment under the high piece rate than under the low piece rate.

To investigate whether written upfront messages have an impact on performance, workers either receive no message, a praise message, or a reference point message. Workers who received a message saw a simple screen before starting to work on the task. The praise message reads as follows: *"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!"* The reference point message is: *"Efficient work is important. Please try to submit at least 25 fragments."* We included the first sentence in the message to provide a mild justification for asking for a specific amount of output and to make the two messages more similar in length. Workers could leave the message screen at any time by clicking on a button to proceed to the work task.

The complete instructions provided to the workers can be found in Section 6.2 of the Appendix.

For the purpose of comparison, we combine the three message settings with each compensation scheme, respectively. The resulting 3x3 treatment design is summarized in Table 1.[2]

[Table 1 about here]

3.2.3 Sample and procedures study 1

For our study 1, we invited a total of 2700 workers from MTurk. 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. For the design and conduct of the study, 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). 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.[3]

Workers who accepted the job offer followed a link to an external website (Qualtrics) that we used for data collection. After workers gave their consent to 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.[4]

[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 study 1. Workers were, on average, 36 years old, possessed a two-year college degree, and were only vaguely familiar with Latin. About five percent

[2] As a robustness check, we conducted treatments where workers receive no message or the praise message with no, low, and high piece rates where we told workers that their work would be approved automatically. We thus cut off any potential concerns that workers may have regarding the impact of their performance in our task on their approval rates. As the results do not differ compared to treatments where we did not explicitly mention automatic approvement, we pooled the data.

[3] We find that 30 workers in our sample restart their work task, which however did not result in any selection effect.

[4] 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 thus had to be excluded after data collection.

used 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 are balanced with respect to all of these characteristics.

Altogether, workers spent on average 13 minutes to complete the work task and the survey. 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 study had been completed. Payments based on performance were transferred within two days after the study was conducted.

3.3 Results study 1

3.3.1 Quantity

We first address the question of whether changes in monetary incentives as well as upfront motivational messages affect output quantity, measured as the number of fragments submitted per worker. In a first step, we focus on differences between distributions of the number of fragments submitted. Figure 2 plots the inverse cumulative distribution function (ICDF) for the number of submitted fragments separated by the type of intervention. In particular, the upper panel of the figure presents the data from all treatments split by the type of monetary incentives provided. Thus, the No-piece-rate ICDF includes all treatments without monetary incentives no matter if an upfront message was used or not. The Low-piece-rate ICDF shows data from all treatments with a low piece rate and the data used for the High-piece-rate ICDF contains all treatments where a high piece rate was offered. The figure allows us to initially study the impact of monetary incentives without taking into account potential interaction effects between monetary incentives and upfront messages.[5]

[Figure 2 about here]

We see from the top panel in Figure 2 that the distribution of the number of submitted fragments including all treatments without a piece rate is stochastically dominated by the distribution of submitted fragments including data from all treatments with a low as well as a high piece rate (two-sided Kolmogorov-Smirnov test (KS), $p = 0.008$ and $p = 0.012$, respectively).[6] In particular, the vertical shift of the ICDF from treatments with a low and a high piece rate indicates that larger monetary rewards appear to increase performance for low and high productivity workers evenly.

The bottom panel in Figure 2 plots the corresponding ICDF dividing the dataset by the use of upfront messages. In this panel, the No-message ICDF contains all treatments without an upfront message no matter if a piece rate was paid or not. The Praise ICDF depicts the data from all treatments where

[5]Figure S2 in the Appendix displays the means and standard deviations for the number of fragments submitted in each treatment.

[6]See Heathcote, Brown, Wagenmakers, and Eidels (2010) for a discussion of stochastic dominance tests.

an upfront praise message was shown whereas the Reference-point ICDF shows all data from treatments including a reference point message. We find that the ICDF for the number of submitted fragments from subjects confronted with a praise message lies below the same function from the treatments where no message was sent (KS, $p = 0.054$). This observation suggests that praising workers prior to work tends to lower overall performance. In contrast, the ICDF from the treatments including a reference point initially dominates the corresponding function from the treatments with no message, whereas being dominated once the reference point is reached. The latter indicates that the explicit setting of a reference point prior to work increases performance below the target output but it decreases performance above it, harmonizing the exerted effort levels of workers. A comparison of variances supports this impression showing that the variance in produced output is significantly lower in the treatments with a reference point than in the treatments with no message and treatments with praise messages (two-sided Levene's variance comparison test, $p < 0.001$ for both comparisons using data of the pooled treatments, respectively).

Next, we estimate the average treatment effect of increasing the piece rate per submitted fragment and the average effect of using praise or reference points on output as well as estimating their interaction effects on output quantity. We do so by using ordinary least squares (OLS) regressions with robust standard errors and employ a series of nested versions for the following difference-in-difference specification for our 3x3 factorial design:

$$\begin{aligned} Y_i = {} & \beta_0 + \beta_1 Low_i + \beta_2 High_i + \beta_3 Praise_i + \beta_4 ReferencePoint_i \\ & + \beta_5 Low_i \times Praise_i + \beta_6 High_i \times Praise_i \\ & + \beta_7 Low_i \times ReferencePoint_i + \beta_8 High_i \times ReferencePoint_i \\ & + \gamma X_i + \varepsilon_i, \end{aligned} \tag{1}$$

where Y captures the number of submitted fragments, Low, $High$, $Praise$, and $ReferencePoint$ are indicator variables for each monetary and non-monetary intervention, X_i is a vector of background characteristics for each worker i and ε_i is an error term. Note that the treatment effects can be captured by a combination of indicator variables. For instance, in the full specification, the indicator variable Low corresponds to the *Low piece rate + No message* treatment effect. The *Low piece rate + Praise* treatment effect can be estimated by adding the coefficients Low, $Praise$, and the interaction between both denoted by $Low \times Praise$ respectively.

Model I in Table 3 reports main effect estimates for increasing piece rates from zero to \$0.01 and \$0.05. Estimate results reveal an increase in average output of 1.40 fragments ($p < 0.001$) for a low piece rate and an increase in average output of 1.42 fragments ($p < 0.001$) for a high piece rate. These changes correspond to a relative increase in average output of about 6.3 percent when compared to using no piece rate payment. Furthermore, we fail to identify any difference in effects between the two piece rates suggesting that the minimum piece rate

payment of one Cent increases worker output as much as a five times higher piece rate ($F\text{-}test,\ diff = 0.02,\ p = 0.963$).

Model I also lists main effect estimates for praising workers or communicating a reference point to them prior to work. We find that communicating reference points to workers insignificantly lowers workers' performance by 0.3 fragments ($p = 0.425$) whereas praising them significantly decreases output by 1.2 fragments ($p = 0.006$) relative to all cases with no upfront motivational message. Additionally, we estimate that the negative effect of praise is significantly stronger than articulating reference points to workers ($diff = 0.882,\ p = 0.031$).

Model II in Table 3 reports both main and interaction effects of all monetary and non-monetary interventions, which enables us to disentangle the effects of our different treatments. We do not find any significant interaction between praising workers prior to work and increased monetary incentives. In particular, whereas the low and high piece rate significantly increase the average number of submitted fragments ($p = 0.013$ and $p = 0.003$, respectively), the *Low piece rate* × *Praise* and *High piece rate* × *Praise* indicator variable estimates remain insignificantly small ($p = 0.802$ *and* $p = 0.884$, respectively), suggesting that praising workers upfront does not curb or amplify monetary incentives. In contrast, we identify from the *High piece rate* × *Reference point* indicator variable estimate that the expression of an explicit reference point curbs the positive effects that result from using a high monetary reward per submitted fragment ($p = 0.031$).[7]

Model III adds a set of worker background variables and shows the robustness of the results discussed so far. Background variables include gender, age, education, device used for the work task, and knowledge of Latin. From the set of background variables, we find that older workers submit, on average, fewer fragments whereas more educated workers and women show a higher work performance in the task. Knowledge of Latin is also predictive for higher worker output in the text transcription task whereas mobile users, on average, submit five fragments fewer than non-mobile device users.

3.3.2 Quality

Next, we assess the quality of each submitted fragment by computing the Levenshtein edit distance to the correct fragment (Levenshtein, 1966). In particular, we calculate 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 and apply a unit cost to each edit operation. We then normalize the processed edit distance by the upper bound of transforming the submitted fragment into the correct fragment obtaining a ratio of dis-similarity of the two fragments that we interpret as the error rate. Workers could use the "?" character as a wildcard if they were unable

[7]The *Low piece rate* × *Reference point* only becomes weakly statistically significant if we add controls in Model III. An F-test for the hypothesis that the coefficients for *High piece rate*, *High piece rate* × *Reference point*, *Low piece rate* and *Low piece rate* × *Reference point* are jointly different from zero cannot be rejected ($p = 0.603$).

to identify the actual character in the presented fragment. We see that workers on average make 0.54 times use of the wildcard. Disregarding the use of the wildcard character when calculating the error rate does not change any result. We report means and standard deviations for the average error rate in Figure S3 in the Appendix.

Following the regression specification in Equation (1), Table 4 presents results from a series of nested random-effects panel regressions, where the dependent variable in each regression captures the error rate of a submitted fragment the worker submitted.[8] Model I reports main effect estimate results of changing the monetary and non-monetary interventions. We find that neither changes in the piece rate nor differences in the upfront praise or reference point 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 rate of about 0.018, i.e., fragments which have a dis-similarity of about 1.8 percent with the correct fragment. Model II and III include interaction terms for both intervention dimensions with and without controls, respectively. Estimation results from these models corroborate that the quality of fragments does not systematically vary across treatments. From the set of background variables, we find that female workers as well as more educated workers submit, on average, fragments with a smaller error rate whereas mobile users deliver fragments that are more error prone.

[Table 4 about here]

On the basis of multitasking theory, a plausible concern in our work setting is that workers who type very fast and submit a large number of fragments deliver low-quality work because they neglect the quality dimension of their task. Figure 3 plots for each worker the number of submitted fragments as a share of the total number of fragments a worker could submit (80 in total) against the average error rate for all submitted fragments by treatment. We consider this share as the completion rate a worker achieves.

[Figure 3 about here]

From the set of sample correlation coefficients that we obtain for each monetary and non-monetary treatment combination, we cannot identify a single significant negative linear relationship between workers' completion rate and the average quality of fragments submitted. In marked contrast to our initial hypothesis, we consistently find that workers who manage to submit a larger number of fragments also submit fragments that are characterized by a lower average error rate.[9]

[8] 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.

[9] Table S1 in the Appendix shows regression results for regressing the averaged error rates 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.

We use a randomized instrumental variables approach (Sajons, 2020) in order to check whether the positive relationship between quality and quantity is merely associational or if an increase in average output indeed comes without the cost of fragment quality. In particular, we employ a two-stage least squares estimation (2SLS) where we treat quantity as the endogenous regressor to predict quality. As instruments for quantity, we use a linear combination of our treatment variables that offer variation in quantity independent of unobserved worker characteristics.

Results of the estimation are shown in Table 5, where we report both OLS and 2SLS estimation results (Antonakis et al., 2010). Model I shows OLS estimation results, which indicate that an increase in a worker's share of fragments submitted has a significant negative effect on the average error rate ($p < 0.001$). The size of the estimate reveals that a one percentage point increase in a worker's completion rate decreases the average error rate by 0.036 percentage points.

Model II and III report the 2SLS results. Model II presents the first- and Model III the second stage estimation results, respectively. First stage results show the positive effect of both low and high piece rates, the negative effect of praise on worker output and the negative interaction of reference points with higher monetary incentives that were previously shown in Section 3.3.1. The partial F statistic shows that our set of instruments is sufficiently correlated with the suspected endogenous regressor ($F(8, 2680) = 30.87, p < 0.001$). Moreover, a Sargan-Hansen test of overidentifying restrictions is not significant ($\chi^2(7) = 8.25, p = 0.311$) giving support to the validity of our instruments.

Results from the second stage show that the estimated effect of quantity on quality is slightly smaller and statistically insignificant when compared to the OLS estimate result. However, the Wu-Hausman ($F(1, 2672) = 0.204, p = 0.652$) and Durbin scores ($\chi^2 = 0.024, p = 0.651$) are statistically insignificant indicating that we cannot reject the hypothesis that the quantity variable is exogenous and that the OLS estimation result we obtain is inconsistent. We therefore conclude that the positive relationship between quality and quantity is not merely associational and that changes in quantity as a result of our interventions do not come at the expense of lower average fragment quality in our study 1.

3.3.3 Supplementary analysis

A possible explanation for the absence of a negative quantity-quality trade-off under performance pay is that workers were concerned about not receiving their piece rate payment if the delivered quality was too low and therefore, in response, typed more carefully than they would in the absence of such concerns. To address this issue, we employed additional clarification treatments where we explicitly informed workers that we would not check the quality of their submitted fragments. 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 receiving 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 non-monetary intervention and two treatments with the low and high piece rate payment scheme in combination with praise for prior to work.[10] 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 4 about here]

Figure 4 plots completion rates of workers against the average error rate for all submitted fragments 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. 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 which is not statistically different to the coefficient for the treatment counterparts in the original study 1 where we did not use any additional clarifying statement ($z = 0.035,\ p = 0.972$).

This result suggests that the absence of a negative quality-quantity trade-off in our original setting is not driven by asymmetric information concerning the implications of low quality work.[11]

3.4 Discussion study 1

For simple work tasks where output quantity can be easily measured, several empirical studies have shown that piece rates lead to higher output than fixed wages (e.g., Lazear, 2000; Shearer, 2004; DellaVigna and Pope, 2018; Antonakis et al., 2019; Meslec et al., 2020). For our task, paying a piece rate also increases output quantity compared to a fixed wage. Our study 1 also shows that the introduction of a very small piece rate works surprisingly well in the context we consider, whereas the marginal effect of increasing the level of monetary incentives is close to zero. This result contrasts with Gneezy and Rustichini's

[10] 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.

[11] 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.

(2000) "Pay enough or don't pay at all" result, and is more in line with DellaVigna and Pope (2018) 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.

Contrary to our hypothesis derived from multitasking theory, we neither find negative effects of piece rates on the quality of work, nor more generally a negative correlation between quantity and quality of output, not even if the employer points out that work quality will neither affect payments nor approval rates. 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. Our results are in contrast to the results of two recent empirical papers that study worker behavior in traditional employment relationships and argue that the absence of multitasking problems under piece rates is due to reputational concerns. Hong, Hossain, List, and Tanaka (2018) 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 employer, which is crucial to fully eliminate reputational concerns of workers. In a similar spirit, Al-Ubaydli, Andersen, Gneezy, and List (2015) propose that workers' uncertainty about the employer's monitoring technology can even lead to higher quality under piece rates than under fixed wages.

Sending a simple message that praises workers for their past performance before the work phase inhibits or even decreases workers' output in our study. This result is puzzling at first sight, but it may show that non-monetary motivational interventions can also have negative performance effects. However, the reduction in performance could also 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 who receive a reference point message to also react negatively to the message because they spend a substantial amount of time reading the message as well.[12] However, we find no indication of a negative effect of our reference point message. Hence, we do not believe that the negative effect of praise on performance is driven by interrupting workers per se but by the content of the message.

Psychologists have studied praise as a social reinforcer and found that praising people can be ineffective or even dysfunctional (Delin and Baumeister, 1994). Baumeister, Hutton, and Cairns (1990) propose three mechanisms that can explain a negative impact of praise on task performance. First, praising may cause people to feel that they no longer need to try hard, leading them to reduce subsequent effort. Second, praise may convey an implicit demand for continued high

[12]Workers who receive a message are presented with a screen displaying the text message prior to work. Figure S1 in the Appendix shows that workers spend on average approximately 6 and 16 seconds reading the reference point message and praise message, respectively. Note that the reference point message is substantially shorter than the praise message.

performance, leading to choking under pressure. Third, receiving praise makes people self-conscious, which impairs their performance. One or a combination of these mechanisms could be at work in our setting.

The short and simple reference point message might not have been strong enough to trigger reference-dependent preferences and might have been perceived rather as a suggestion than a formal request to achieve this reference point. Our reference point resembles an externally assigned goal. Psychologists assert that assigned goals can be effective, but more so if the goal is ambitious and the assigning person explains the goal and expresses confidence that the goal can be achieved (see, e.g., Locke and Latham (2002) for an overview of the literature). Our simple message did not carry any such information. Thus, the treatment manipulation might have been too subtle to trigger higher performance. However, our data also revealed that the variance in the treatments with reference points is lower than in the treatments using praise or no upfront message, suggesting that workers have to some extent reacted to our intervention. The effects for high and low performing workers might have cancelled each other out. In addition, the interaction effects with performance pay indicate that the positive effect of paying a piece rate on average performance is likely to be offset by the introduction of a reference point. An intervention that properly incorporates the recommendations of the goal setting literature might enhance average performance, e.g., by triggering higher output from low performers or diminishing the output decline of high performers relative to a less sophisticated goal intervention. Overall, the negative effect of our praise intervention calls for further analysis to clarify whether other and stronger forms of non-monetary interventions will also backfire in an online setting. We, therefore, conducted a second study where we address the points raised above.

4 Study 2

4.1 Aim and hypotheses study 2

In study 2, we focus on non-monetary interventions based on charismatic leadership communication techniques. In particular, we investigate whether and how communication tactics used by charismatic leaders affect the performance of online workers, building on the concept of charismatic leadership as defined by Antonakis et al. (2011, 2016, 2019). Antonakis et al. (2016) define charismatic leadership as "values-based, symbolic, and emotion-laden signaling," which provides us with a suitable definition and operationalization of communication tools to develop an experimental design. Charismatic leaders use communication tactics, which can be organized in three major categories that can be reliably coded. The first category is "frame and vision" by which the leader tries to draw attention to the key issues of the job. The second category is "substance" which is used to justify the mission and announce strategic goals. "Frame and vision" can be provided by (i) metaphors, (ii) rhetorical questions, (iii) stories and anecdotes, (vi) contrasts, and (v) three-part lists. "Substance" can be induced by (vi)

expressing moral conviction, (vii) expressing sentiments of the collective, (viii) setting high and ambitious goals, and (ix) creating the confidence that workers will be able to reach these goals. The two categories "frame and vision" and "substance" rely on verbal communication tactics, whereas the third category "delivery" is triggered by non-verbal tactics. By the use of voice, body gestures, and facial expressions the leader can demonstrate passion and confidence. As we want to study the effects of written messages in online labor markets, we do not implement the third category and thus focus on the first two.

Similar to study 1, we employed a transcription task to measure both quality and quantity of the submitted work. Our main research question is if verbal tactics providing "frame and vision" as well as "substance" in a purely written form will be sufficient to increase output. In addition, we are interested in disentangling the effects of goal setting from other verbal CLTs, as simple quantitative goals are often used in isolation in practice and also have been studied in isolation before.

We conduct four treatments, named *Neutral*, *Goal*, *Charisma without goal* and *Full charisma* that differ in the CLTs employed. In the *Neutral* treatment, the task is explained as neutrally as possible. The *Goal* treatment sets a specific quantitative goal utilizing the verbal CLTs (viii) and (ix). By contrast, *Charisma without goal* makes use of the remaining CLTs (only contrasts are not used) without setting a quantitative goal. Finally, *Full charisma* combines all CLTs used in the former two treatments. Thus, *Charisma without goal* features fewer elements triggering "substance" in comparison to the *Full charisma* intervention. The *Goal* treatment, in contrast, focuses only on a subset of CLTs related to the "substance" category.

For the derivation of our hypotheses, we build on the theoretical economic framework proposed by Antonakis et al. (2019). They assume that workers receive positive intrinsic utility from working on their task, and that the absolute and the marginal intrinsic utility from working increases in the perceived charisma of the leader, without addressing the specific psychological mechanism through which charisma impacts utility. Accordingly, if workers perceive the leader as more charismatic, they will work harder. Based on this framework, we expect that both quantity and quality of work increase if workers perceive the leader as more charismatic.[13] The question, however, is if our interventions are sufficient to trigger higher perceived charisma. We expect that the CLTs employed in the *Full charisma* treatment will increase the perceived charisma and therefore lead to a significantly higher performance than the *Neutral* treatment. It is, however, unclear how the use of only subsets of verbal CLTs are perceived. In particular, in our setting, we cut off important non-verbal channels that a leader can typically use. It may be that subsets of CLTs are too weak to increase the perception of charisma, but it is also possible that they are effective. We, therefore, expect to find either higher performance or no difference in performances when comparing the *Charisma without goal* with the *Neutral* treat-

[13]This assumption is also driven by the fact that we did not find a negative relationship between quantity and quality of work in our first study.

ment. The *Goal* treatment also employs only a subset of CLTs. Nevertheless, we expect these CLTs to increase performance relative to the *Neutral* treatment because research in psychology (e.g., Locke and Latham, 2002) and economics (e.g., Corgnet et al., 2015) asserts that assigning goals increases performance.

4.2 Design study 2

4.2.1 Work task study 2

The workers transcribed historic documents from the Frick Collection and Frick Art Reference Library Archives. All documents are typed letters written in English. The transcribed documents will become part of the collection and will be accessible and searchable by researchers and the general public.[14] The task, therefore, has a clear meaning and adds value and at the same time also requires effort and attention to detail. We divided all letters into fragments and constructed 15 batches of fragment groups. Each batch consists of a sequence of fragments, where the length of a fragment (i.e., its number of characters) at a given position in the sequence is roughly constant across all batches. In each treatment workers were randomly assigned to one of the batches. We let 20 workers work on the same batch to provide us with sufficient data for quality control. As in study 1, workers could type fragments for a total duration of ten minutes. They received one fragment at a time on screen and got a new fragment after each submission.

4.2.2 Treatments study 2

We again use a between-subject design to systematically investigate the impact of motivational techniques, in particular, charismatic leadership tactics, including quantitative output goals, on worker performance. We conducted four treatments labeled *Neutral*, *Goal*, *Charisma without goal*, and *Full charisma* which differ by the written instructions workers receive prior to work. All treatment instructions contained the same information about the nature of the task and are of similar length. Workers in the *Neutral* treatment received standard instructions informing them about the purpose of their work, the collaboration with the Frick Collection and Frick Art Reference Library Archives, and that they would be working together with other workers to preserve historic documents. The complete instructions for each treatment can be found in Section 6.3 of the Appendix.

The *Charisma without goal* treatment differed from the *Neutral* treatment only in that instructions have been written in a more charismatic way using nonverbal charismatic leadership tactics (CLT) according to Antonakis et al. (2011); Antonakis, Fenley, and Liechti (2012) wherever possible. Inspired by Antonakis et al. (2019), in particular we employed metaphors, rhetorical questions, stories, three-part-lists, moral conviction, and raised sentiments of the collective. We

[14]We are grateful that the Frick Collection and Frick Art Reference Library Archives have agreed to collaborate with us.

did not use goal-related CLTs or explicitly state a quantitative goal in this treatment. The *Goal* treatment is identical to the *Neutral* treatment but contains an additional paragraph where we communicate a quantitative output goal. In particular, we add the two CLTs "setting high and ambitious goals" and "creating confidence that the goal can be achieved" to the instructions of the *Neutral* treatment. We provide workers with the additional information that workers in similar HITs previously managed to transcribe 25 fragments on average and we ask them to score at least 34 fragments.[15] Moreover, we clarify that scoring 34 fragments was a challenging yet achievable goal. We also told workers that we were confident they would reach their goal because of their work experience. The *Full charisma* treatment combines the non-goal related charismatic leadership tactics from the *Charisma without goal* treatment and the goal related CLTS from the *Goal* treatment in the instructions. The *Full charisma* treatment therefore, contains the most CLTs from all our treatments and triggers "frame and vision" as well as "substance." The resulting 2x2 treatment design is summarized in Table 6.

[Table 6 about here]

4.2.3 Sample and procedures study 2

A total of 1800 workers participated in our second study. We posted the same job advertisement on MTurk for a data entry task that we used in study 1 with the only difference being that we raised the fixed payment to 3 USD. Our selection criteria (98% approval rate or higher, at least 500 previously approved HITs, location U.S.) remain also unchanged. We excluded workers that have participated in previous sessions of study 1 or the pilot study we used to determine the goal for our Goal treatment.[16] After workers have accepted the HIT, they followed a link to Qualtrics which hosts our study. All workers were randomly allocated to treatments and could work on the task for ten minutes. Afterwards they completed a short survey containing questions on demographics, information on the device used for working on the study, touch typing ability, familiarity with Frick Collection and Frick Art Reference Library Archives, questions on motivation, identification with the mission to preserve historic documents, and the impact of the Covid-19 pandemic on work life.

[Table 7 about here]

Table 7 provides an overview of the background characteristics of subjects participating in study 2. Workers were, on average, 38 years old and possessed

[15]The goal was determined based on a pilot study with 120 workers who worked on the *Neutral* treatment. Reaching the goal translates to being among the top 28 percent performers in the pilot study.

[16]One worker accepted the invitation but never actually worked on the task and is therefore missing from the sample. In addition, the timer of the work task did not work properly for 31 workers who had to be excluded after data collection. In addition, we also omit 10 workers from the analysis who self-identify as neither male or female.

a four year college degree. About four percent used 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. Figure S4 in the Appendix also shows that workers spend, on average, the same amount of time reading the intervention messages they receive before beginning to work.

4.2.4 Manipulation check study 2

The objective manipulation check was executed by external evaluators in order to avoid that workers become aware of the different instructions and might adjust their behavior (see Lonati, Quiroga, Zehnder, and Antonakis, 2018). The check was done by two trained research assistants who independently coded each treatment instruction. The research assistants were instructed to mark the occurrence of the verbal charismatic leadership tactics on sentence level (see Tables S6, S7, S8 in the Appendix). We calculated the intercoder reliability for each treatment. For the *Neutral* treatment (n=15 sentences) the coders agreed on 99.26% of the coding events (15 sentences x 9 charismatic leadership tactics). The agreement level can be tested against chance agreement using Cohen's kappa with $\kappa = 0.85$, $se = 0.085$, $z = 10.02$ and $p < 0.01$ revealing a substantial or almost perfect alignment of coders for all treatment interventions (see Landis and Koch, 1977). The agreement percentage is 96.73% for the *Charisma without goal* treatment (n=17 sentences with $\kappa = 0.74$, $se = 0.081$, $z = 9.22$ and $p < 0.01$)), 98.15% for the *Goal* treatment (n=18 sentences with $\kappa = 0.76$, $se = 0.078$, $z = 9.70$ and $p < 0.01$)) and 96.11% for the *Full charisma* treatment (n=20 sentences with $\kappa = 0.72$, $se = 0.075$, $z = 9.66$ and $p < 0.01$)). The agreement rates are rather similar for each treatment and both coders reconciled their coding after having coded individually until they reached an agreement.

4.3 Results study 2

4.3.1 Quantity

For analysing the performance of our workers, we again first focus on differences between distributions of the number of fragments submitted. Figure 6 plots the inverse cumulative distribution function (ICDF) for the number of submitted fragments for each treatment. We find that the ICDF from the *Full charisma* treatment, where we combine goal-related and non-goal related CLTs, lies furthest to the right of all ICDFs and stochastically dominates the distribution of submitted fragments relative to all other treatments we employ (KS, two sided, vs. *Charisma without goal*: $p < 0.001$, vs. *Goal*: $p = 0.028$ and vs. *Neutral*: $p = 0.033$, respectively). In comparison, the ICDF from the *Charisma without goal* treatment lies furthest to the left and is also stochastically dominated by the *Goal* treatment (KS, two sided, $p = 0.046$). This observation suggests that the isolated use of non goal-related CLTs yields the poorest performance among all

treatments in terms of output produced whereas the combination of goal-related and non-goal related CLTs results in the highest.[17]

[Figure 6 about here]

Similar to study 1, we estimate the average treatment effects of our interventions by using ordinary least squares (OLS) regressions with robust standard errors and employ a series of nested versions for the following difference-in-difference specification for our 2x2 factorial design:

$$Y_i = \beta_0 + \beta_1 Goal_i + \beta_2 Charisma_i + \beta_3 Goal_i \times Charisma_i + \gamma X_i + \varepsilon_i \quad (2)$$

where Y captures the number of submitted fragments and *Goal* is an indicator variable for treatments that employ goal-related CLTs including a quantitative goal. $Charisma$ is an indicator variable for the use of non-goal related CLTs and X_i is a vector of background characteristics for each worker i and ε_i is an error term. In the full specification, the indicator variable $Charisma$ corresponds to the *Charisma without goal* treatment whereas the *Full charisma* treatment effects can be estimated by adding the coefficients of $Goal$, $Charisma$, and the interaction between both denoted by $Goal \times Charisma$.

Model I in Table 8 reports main effect estimates result which can be interpreted as measuring the overall difference in average output for both treatments that make use of goal-related CLTs (*Goal* and *Full charisma* treatment) against not using any goal related CLTs (*Neutral* treatment) by the *Goal* coefficient. The $Charisma$ coefficient reveals the difference in average output of using non goal-related CLTs and a broad set of CLTs (*Charisma without goal* and *Full charisma* treatment) against not using any CLTs at all (*Neutral* treatment). At first sight, we see a statistically significant increase in average output of 1.72 fragments ($p < 0.001$) when we communicate output goals and use goal-related CLTs to workers prior to work and an overall positive, albeit insignificant effect of using non-goal related and using a broad set of CLTs ($p = 0.624$).

Model II adds interaction effects to the estimation allowing us to disentangle the effects of our four treatments and to study the impact of using a broad set of CLTs including goals. We find a significant interaction between using non-goal related CLTs and goal related CLTs indicating that the usage of a broad set of CLTs is needed in online labor markets. In particular, setting output goals and using goal related CLTs alone has no significant effect on average performance ($p = 0.695$) rendering our *Goal* treatment intervention ineffective to enhance performance. Surprisingly, employing charismatic leadership tactics without the inclusion of goal related CLTs even significantly decrease average performance by 1.75 fragments ($p = 0.022$) and backfires in our *Charisma without goal* treatment. In contrast, using the broader set of charismatic leadership tactics by combining our two CLT interventions significantly increases average output per worker ($p < 0.001$). Specifically, the *Charisma* $\times$ *Goal* indicator

[17] We find no difference between the ICDF from the *Goal* and *Neutral* treatment ($p = 0.901$).

variable estimate indicates that the difference-in-difference effect size, i.e. the difference in average output of using goal related CLTs to not using them is about 4.06 fragments larger when combining them with non-goal related CLTs. The latter combination yields the highest average output of all interventions of 31.3 fragments per worker. We report means and standard deviations for the number of fragments submitted in Figure S5 in the Appendix.

Model III and IV add a set of worker background variables and fragment group specific intercepts and show that results remain robust to the inclusion of these variables. Background variables include gender, age, education, device used for the work task. From the set of background variables, we find that older workers submit, on average, fewer fragments whereas more educated workers and females show a higher work performance in the task. Using a mobile device in the text transcription task decreases work performance significantly.

4.3.2 Quality

For assessing the quality of each submitted fragment, we compute the Levenshtein edit distance to the correct fragment and then normalize the distance to obtaining an error rate per fragment.[18] Following the regression specification in Equation (2), Table 9 presents results from a series of nested random-effects panel regressions, where the dependent variable in each regression captures the error rate of a submitted fragment the worker submitted.[19] Model I reports main effect estimate results of using non-goal related CLTs and using only goal related CLTs whereas Model II, III, and IV also include interaction terms for both interventions as well as background characteristics of workers as controls and intercepts for different fragment groups, respectively. Overall, we find that workers submit, on average and over time fragments that have an average error rate of about 0.021 ($p < 0.001$) independent of their treatment affiliation. None of our treatment interventions has any statistically significant effect on the average quality of work. From the set of background variables, we find that female workers submit, on average, fragments with a smaller error rate whereas mobile users deliver fragments that are significantly more error prone.

[Table 9 about here]

We again investigate whether workers who submit a larger number of fragments deliver low-quality work because they neglect the quality dimension of their task. Figure 7 plots the completion rate for each worker, i.e. the number of submitted fragments as a share of the total number of fragments a worker could submit (110 in total), against the average error rate for all submitted fragments by treatment.

[18] We compare the entered output of the workers to determine the correct spelling of the fragment. If we have only one observation for a fragment, we let a research assistant type the fragment as well to allow us to control for quality.

[19] We report means and standard deviations for the average error rate in Figure S6.

[Figure 7 about here]

We fail to identify any significant negative linear relationship between the share of submitted fragments a worker submits and their average quality but instead find for all treatments that workers who produce more fragments also submit fragments that are characterized by a lower average error rate.[20]

We make use of a randomized instrumental variables approach to test whether average fragment quality is unaffected by changes in average output as a result of our treatment interventions or whether the relationship depicted in Figure 7 is purely associational. In particular, we employ the same 2SLS estimation procedure and treat quantity as the endogenous regressor to predict quality. Our treatment variables serve again as our instruments.

Results of the estimation are shown in Table 10. Model I shows OLS estimation results which indicate that an increase in the completion rate has a significant negative effect on the average error rate ($p < 0.001$). The size of the estimate reveals that a one percentage point increase in a worker's completion rate decreases the average error rate by 0.089 percentage points.

Model II and III report the 2SLS results with Model II presenting the first- and model III the second stage estimation results, respectively. First stage results show the negative effect of non-goal related CLTs and using the broad set of CLTs including goals. The partial F-statistic shows that our set of instruments is sufficiently correlated with the suspected endogenous regressor ($F(3, 1758) = 25.2, p < .001$). Moreover, a Sargan-Hansen test of overidentifying restrictions is not significant ($\chi^2(2) = 0.110, p = 0.946$) giving support to the validity of our instruments.

Results from the second stage show that estimated effect of quantity on quality is slightly smaller and statistically insignificant when compared to the OLS estimate result. However, the Wu-Hausman ($F(1, 1751) = 0.09, p = 0.765$) and Durbin scores ($\chi^2 = 0.09, p = 0.764$) are statistically insignificant indicating that we cannot reject the hypothesis that the quantity variable is exogenous and that the OLS estimation result we obtain is inconsistent. We therefore conclude that the positive relationship between quality and quantity is not merely associational and that changes in quantity as a result of our interventions do not come at the expense of average fragment quality in our study 2.

4.4 Discussion study 2

In study 2, we concentrated on communication techniques used by charismatic leaders to investigate if the effects of the non-monetary interventions in our first study can be replicated or if verbal charismatic leadership techniques can trigger higher performance in an online labor market.

[20]Table S9 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.

Our first study has shown that providing a reference point for output quantity had no significant effects on output quantity or quality. This result is backed up by our second study where our goal intervention did not significantly affect performance either. We again formulated a quantitative goal but also provided substance by justifying the chosen goal and expressing confidence that the goal is challenging but achievable for the workers. We stated clearly on what we based the goal, namely on the performance of other workers in previous sessions. In addition, we provided information about the average performance of other workers (25 fragments) to set a challenging reference point at 34 fragments. To raise workers' self-efficacy belief, we also expressed our confidence that they will be able to reach the goal. Nevertheless, we did not find any performance effects, which indicates that a goal related subset of CLTs that aim at providing only "substance" but cannot address "frame and vision" is not sufficient to trigger the perception of charisma in our setting. In contrast to our results on the reference point message in study 1, we did not find evidence for an effort-harmonizing effect of the goal message in study 2.

The *Charisma without goal* treatment encompasses a subset of CLTs aimed at providing "frame and vision" and to a lesser extent "substance" because we did not use goal related CLTs in this treatment. We expected this intervention to have a positive or neutral impact on both performance dimensions. However, the results show that this intervention leads to lower output than the *Neutral* treatment and thus backfires. We observed a similar pattern in the *Praise* treatment of study 1. We conclude that, when we only employ a subset of CLTs such that the category "substance" is underrepresented, workers perceive our written instructions not as intended. Using such a subset of CLTs even turns out to be harmful for the employer because it reduces quantity considerably. We can only speculate about the underlying reasons for this finding. Maybe workers perceive an intervention that lacks important elements of substance in the form of specific goals as not authentic and artificial, in particular if the leader communicates only in writing. This possible explanation is supported by our findings in the *Full charisma* treatment.

We implement all CLTs from the *Charisma without goal* and the *Goal* treatments in the *Full charisma* treatment making use of a broad set of charismatic communication tactics triggering both categories, "frame and vision" as well as "substance." This intervention leads to a considerable increase in output quantity and we observe a strong complementarity between goal related CLTs and other verbal CLTs. Our results indicate that, in an online setting, it is important to use a set of CLTs that covers both verbal categories of charismatic leadership as broadly as possible. Quantitative goals may provide focus and align effort of the workers towards a common target, but we also need to provide "frame and vision" to raise the workers' attention and add more substance to the message by using for instance sentiments of the collective or moral convictions to justify the mission. Using only "frame and vision" oriented CLTs can dramatically backfire whereas the use of only goal-related CLTs seems to be too weak to raise output levels. However, the performance-enhancing effect of pure verbal CLTs

in an online labor market setting with written communication only is impressive and shows the power of well-balanced communication even in the absence of non-verbal clues.

5 General discussion

The results of our two studies have provided rich insights into the potential to motivate workers on online labors markets with either transactional techniques, in particular monetary incentives, or by applying communication techniques, in particular short upfront messages and charismatic leadership tactics. In the following, we discuss some general findings and limitations of our two studies.

As already discussed, monetary incentives work but a higher piece rate does not lead to higher output in our set-up. This result is important because, holding fixed payments constant, higher monetary incentives entail higher costs of labor and higher payments for each delivered unit. Employers are therefore better off if they implement a rather low piece rate, at least in our setting. However, we offer a rather generous fixed payment to our workers in both studies which might have already motivated workers to some extent so that performance pay was rendered less effective. Future studies should explore if lower fixed payments or even pure performance pay lead to similar performance patterns.

Praising workers for their past good performance or using only a subset of non-goal related verbal CLTs backfired in our studies, whereas the usage of a broad CLT set led to a substantial increase in delivered output. Our findings indicate that employers in online labor markets need to pay attention to what and how they communicate. An unreflected usage of praise or non-goal related subsets of CLTs can be even harmful and result in lower performance. In order to evoke the positive effect associated with charismatic communication, employers need to use a broad set of CLTs addressing both categories, "frame and vision" as well as "substance" properly. Our study reveals that goal-related CLTs and non-goal related verbal CLTs are complements and work well even in an online setting with written messages only. The test if and how other combinations of verbal CLTs work in the online labor market has to be left for future studies.

Overall, the strong impact of charismatic communication on output levels in an online labor market with a spot contract and purely written one-way communication from employer to worker is impressive. Usually non-verbal clues contribute highly to being perceived as a charismatic leader and the usage of verbal and non-verbal techniques of charisma correlate quite strongly (Antonakis et al., 2011). However, in our online setting where written upfront communication is usually the only channel of interaction between employers and workers, the use of non-verbal techniques does not seem necessary to achieve a value-based, symbolic, and emotion-laden signaling by the employer. The reason might be that online workers do not expect such non-verbal clues. Our results, therefore, might not be generalizable to traditional work settings where employers have access to richer forms of communication such as audio, video, or face-to-face interaction.

6 Conclusion

In contrast to employees within traditional firms, workers in online labor markets usually work on their own and have no face-to-face contact with employers and coworkers. Communication between worker and employer is typically one-sided and delivered in written instructions of the work task before the worker starts with the assigned task. This setting makes motivating online workers more challenging than motivating workers in traditional employment relationships, which typically entail frequent face-to-face interactions. In this paper we have presented results from two large scale experiments on Amazon Mechanical Turk, investigating the effect of piece rates and different forms of written communication on quality and quantity of the delivered work.

Our results reveal that monetary incentives work in online labor markets but there is no positive relationship between higher piece rates and output levels. Upfront motivational messages in form of praise or only subsets of non-goal related verbal CLTs backfire and lead to a significant reduction in output, whereas the provision of goals and reference points do not lead to significant changes in average performance. Importantly, we find that using a broad set of verbal charismatic leadership tactics including goal-related CLTs enhances performance significantly even though non-verbal tactics such as facial expressions, body language and animated voice are completely missing in our written set-up.

We do not find significant changes in the delivered quality in any of our treatment interventions. Thus, there is no evidence of a multitasking problem in our online labor market setting. Given that the employers usually rely on the workers to deliver high quality, this finding is very promising for posting work on online labor markets. Indeed, we observe a positive correlation between quantity and quality indicating that triggering higher quantitative output is also beneficial in terms of overall work achievements.

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, Nicolas Bastardoz, Philippe Jacquart, and Boas Shamir (2016). "Charisma: An ill-defined and ill-measured gift," *Annual Review of Organizational Psychology and Organizational Behavior*, 3: 293–319.

Antonakis, John, Samuel Bendahan, Philippe Jacquart, and Rafael Lalive (2010). "On making causal claims: A review and recommendations," *The leadership quarterly*, 21(6): 1086–1120.

Antonakis, John, Giovanna d'Adda, Roberto Weber, and Christian Zehnder (2019). "Just words? Just speeches? On the economic value of charismatic leadership," *working paper*.

Antonakis, John, Marika Fenley, and Sue Liechti (2011). "Can charisma be taught? Tests of two interventions," *Academy of Management Learning & Education*, 10(3): 374–396.

Antonakis, John, Marika Fenley, and Sue Liechti (2012). "Learning charisma. transform yourself into the person others want to follow." *Harvard business review*, 90(6): 127–30.

Avolio, Bruce J, Surinder Kahai, and George E Dodge (2000). "E-leadership: Implications for theory, research, and practice," *The leadership quarterly*, 11(4): 615–668.

Baumeister, Roy F., Debra G. Hutton, and Kenneth J. Cairns (1990). "Negative effects of praise on skilled performance," *Basic and applied social psychology*, 11(2): 131–148.

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.

Coase, Ronald (1937). "The nature of the firm," *Economica*, 4(16): 386–405.

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, Edward L. (1971). "Effects of externally mediated rewards on intrinsic motivation," *Journal of Personality and Social Psychology*, 18(1): 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 (2018). "What motivates effort? Evidence and expert forecasts," *The Review of Economic Studies*, 85(2): 1029–1069.

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

Falk, Armin and Urs Fischbacher (2006). "A theory of reciprocity," *Games and economic behavior*, 54(2): 293–315.

Farrell, Anne M., Jonathan H. Grenier, and Justin Leiby (2017). "Scoundrels or stars? theory and evidence on the quality of workers in online labor markets," *The Accounting Review*, 92(1): 93–114.

Gallup (2018). "The gig economy and alternative work arrangements," `https://www.gallup.com/workplace/240878/gig-economy-paper-2018.aspx`.

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.

Heathcote, Andrew, Scott Brown, EJ Wagenmakers, and Ami Eidels (2010). "Distribution-free tests of stochastic dominance for small samples," *Journal of Mathematical Psychology*, 54(5): 454–463.

Hermalin, Benjamin E (1998). "Toward an economic theory of leadership: Leading by example," *American Economic Review*: 1188–1206.

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

Hong, Fuhai, Tanjim Hossain, John A. List, and Migiwa Tanaka (2018). "Testing the theory of multitasking: Evidence from a natural field experiment in Chinese factories," *International Economic Review*, 59(2): 511–536.

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.

ILO (2018). "Digital labour platforms and the future of work: Towards decent work in the online world," `https://www.ilo.org/wcmsp5/groups/public/---dgreports/---dcomm/---publ/documents/publication/wcms_645337.pdf`.

Kässi, Otto and Vili Lehdonvirta (2018). "Online labour index: Measuring the online gig economy for policy and research," *Technological Forecasting and Social Change*, 137: 241–248.

Kosfeld, Michael, Susanne Neckermann, and Xiaolan Yang (2017). "The effects of financial and recognition incentives across work contexts: The role of meaning," *Economic Inquiry*, 55(1): 237–247.

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.

Landis, J. Richard and Gary G. Koch (1977). "The measurement of observer agreement for categorical data," *Biometrics*, 33(1): 159–174.

Lazear, Edward P. (2000). "Performance pay and productivity," *American Economic Review*, 90(5): 1346–1361.

Levenshtein, Vladimir 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, Edwin A. and Gary P. Latham (1984). *Goal Setting: A Motivational Technique That Works*, Englewood Cliffs, NJ:Prentice-Hall.

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

Lonati, Sirio, Bernardo F. Quiroga, Christian Zehnder, and John Antonakis (2018). "On doing relevant and rigorous experiments: Review and recommendations," *Journal of Operations Management*, 64: 19–40.

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

Meslec, Nicoleta, Petru L. Curseu, Oana C. Fodor, and Renata Kenda (2020). "Effects of charismatic leadership and rewards on individual performance," *The Leadership Quarterly*: 101423.

Munger, Michael (2015). "Coase and the sharing economy," in Cento Veljanovski (ed.), "Forever Contemporary: The economics of Ronald Coase," The Institute of Economic Affairs, pp. 187–208.

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, Kathrin (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 (2018). "Your loss is my gain: A recruitment experiment with framed incentives," *Journal of the European Economic Association*, 16: 522–559.

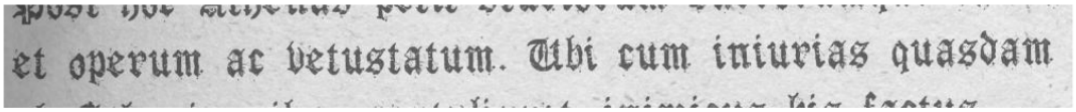
Sajons, Gwendolin B. (2020). "Estimating the causal effect of measured endogenous variables: A tutorial on experimentally randomized instrumental variables," *The Leadership Quarterly*: 101348.

Shearer, Bruce (2004). "Piece rates, fixed wages and incentives: Evidence from a field experiment," *The Review of Economic Studies*, 71(2): 513–534.

The Economist (2020). "What will be the new normal for offices?" `https://www.economist.com/britain/2020/05/09/what-will-be-the-new-normal-for-offices`.

Zehnder, Christian, Holger Herz, and Jean-Philippe Bonardi (2017). "A productive clash of cultures: Injecting economics into leadership research," *The Leadership Quarterly*, 28(1): 65 – 85.

Figure 1: Screenshot of the work task, study 1



Note: The picture shows an example fragment from the task used in study 1 that workers had to transcribe.

Table 1: Treatment table, study 1

	Non-monetary intervention			
Performance pay	No message	Praise	Reference point	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

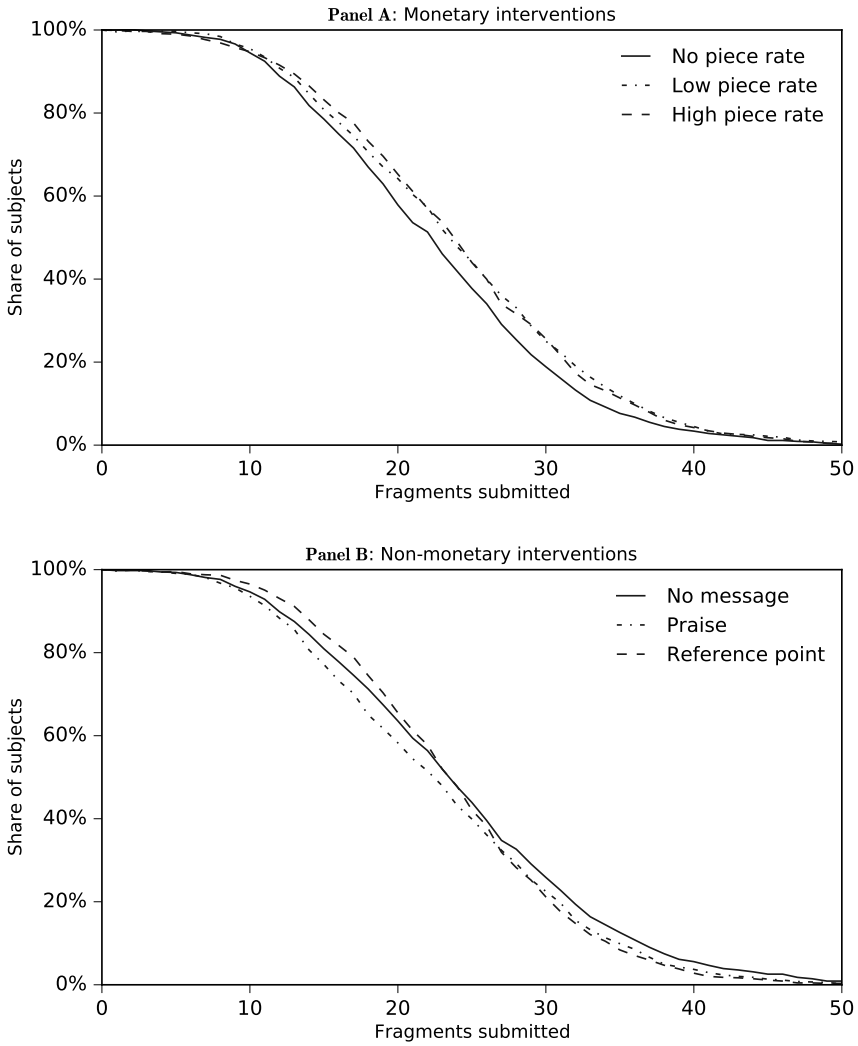
Note: The table gives an overview of the experimental design of study 1 and shows the combination of the monetary and non-monetary treatment interventions. The number of subjects for each treatment cell is indicated as well.

Table 2: Background characteristics of subjects, study 1

		Age	Female	Education	Mobile device	Latin	
Performance pay	**Non-monetary intervention**	**Mean (se)**	**Mean (se)**	**Mean (se)**	**Mean (se)**	**Mean (se)**	**N**
No piece rate	No message	36.28 (0.59)	0.50 (0.03)	3.12 (0.08)	0.05 (0.01)	1.42 (0.04)	300
	Praise	36.04 (0.62)	0.50 (0.03)	3.24 (0.08)	0.03 (0.01)	1.38 (0.04)	292
	Reference point	35.77 (0.65)	0.54 (0.03)	3.12 (0.07)	0.07 (0.01)	1.44 (0.04)	299
Low piece rate	No message	35.87 (0.64)	0.50 (0.03)	3.08 (0.07)	0.07 (0.01)	1.41 (0.04)	295
	Praise	34.49 (0.56)	0.50 (0.03)	3.07 (0.08)	0.04 (0.01)	1.41 (0.04)	301
	Reference point	35.42 (0.64)	0.49 (0.03)	3.15 (0.08)	0.03 (0.01)	1.45 (0.05)	295
High piece rate	No message	34.93 (0.61)	0.46 (0.03)	3.02 (0.08)	0.05 (0.01)	1.46 (0.04)	302
	Praise	35.15 (0.64)	0.52 (0.03)	3.13 (0.07)	0.06 (0.01)	1.40 (0.04)	297
	Reference point	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
Pr(>F)		0.405	0.637	0.740	0.137	0.850	

Note: The table reports background characteristics of subjects participating in study 1. Subjects were recruited through the Amazon Mechanical Turk crowdsourcing 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. Reported are also p-values for the overall regression F-statistic from models in which the respective background characteristic is regressed on all treatment indicator variables.

Figure 2: Fragments submitted, study 1



Note: The figure plots the inverse cumulative distribution function for the number of submitted fragments in all treatments with no, a low, or a high piece rate (Panel A) and all treatments with no message, a praise message, or a reference point message (Panel B).

Table 3: Treatment effects on quantity, study 1

Model	I	II	III
Low piece rate	1.398***	1.950**	1.978***
	(0.422)	(0.780)	(0.748)
High piece rate	1.418***	2.190***	1.968***
	(0.415)	(0.737)	(0.718)
Praise	-1.215***	-1.075	-1.260*
	(0.438)	(0.730)	(0.703)
Reference point	-0.332	0.847	0.781
	(0.417)	(0.690)	(0.669)
Low piece rate		-0.272	-0.474
× Praise		(1.082)	(1.041)
High piece rate		-0.153	0.092
× Praise		(1.045)	(1.015)
Low piece rate		-1.379	-1.610*
× Reference point		(1.020)	(0.974)
High piece rate		-2.160**	-1.960**
× Reference point		(0.999)	(0.968)
Age			-0.193***
			(0.015)
Female			0.728**
			(0.334)
Education			0.529***
			(0.131)
Mobile device			-5.022***
			(0.863)
Latin			0.960**
			(0.374)
Constant	22.718***	22.277***	27.225***
	(0.384)	(0.505)	(0.831)
N	2680	2680	2680
R^2	0.009	0.011	0.084
F	5.752	3.458	18.757
Pr(>F)	0.000	0.001	0.000

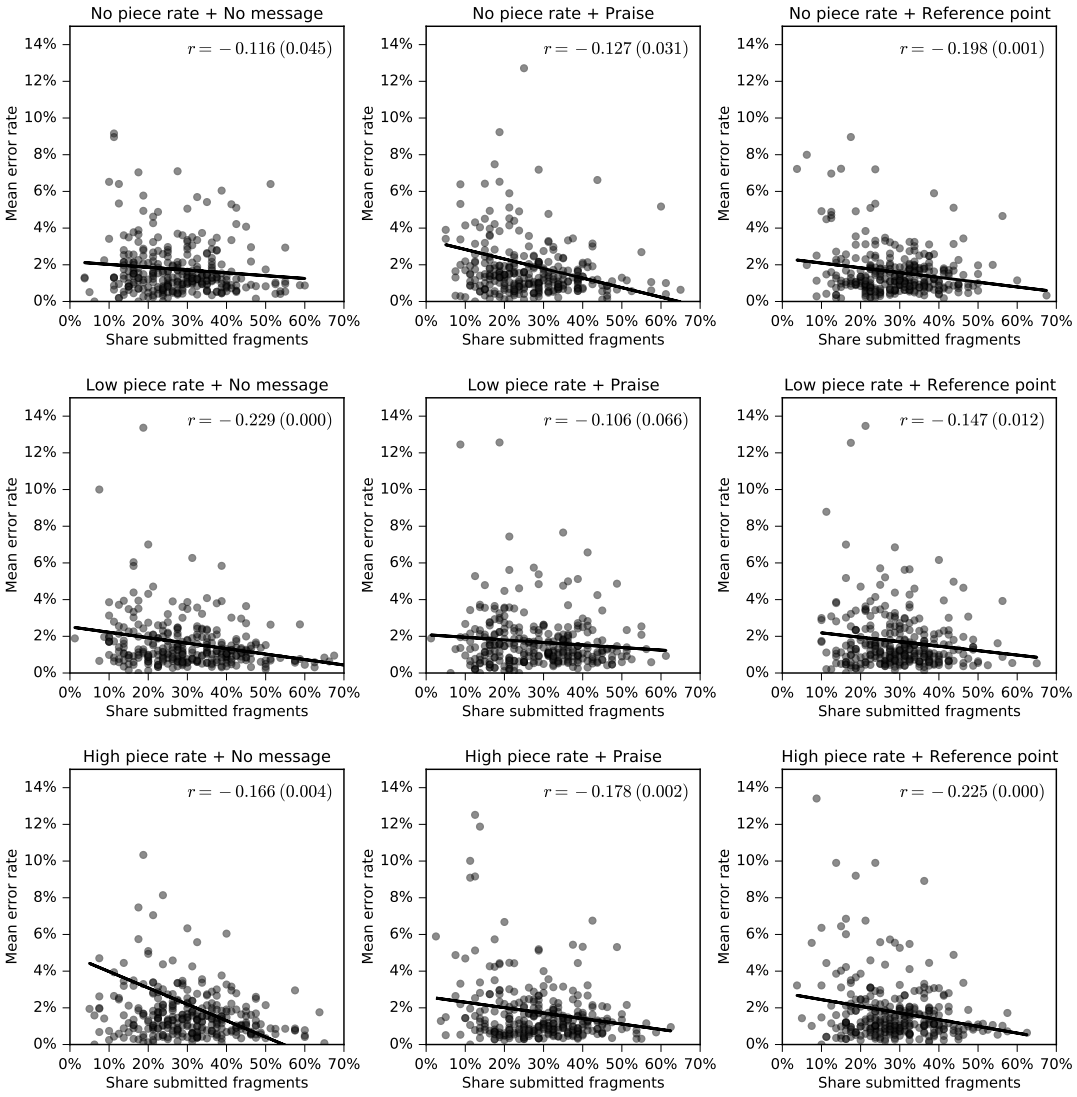
Note: The table reports estimation results for regressions in which the number of fragments submitted per worker is regressed on a set of explanatory variables. "Low piece rate": indicator variable taking the value of one if the treatment used a low piece rate. "High piece rate": indicator variable taking the value of one if the treatment used a high piece rate. "Praise": indicator variable taking the value of one if the treatment praised workers. "Reference point": indicator variable taking the value of one if the treatment set a reference point. "Age": continuous variable measuring a worker's age. "Female": indicator variable taking the value one if the worker is a female. "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": indicator variable taking the value one if the worker used a mobile device. "Latin": indicator variable variable taking the value of one if the worker has at least some knowledge of Latin. Robust standard errors in parentheses ($* : p < 0.1$, $** : p < 0.05$, $*** : p < 0.01$).

Table 4: Treatment effects on quality, study 1

Model	I	II	III
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)
Praise	-0.000	0.002	0.002
	(0.002)	(0.003)	(0.003)
Reference point	-0.001	-0.002	-0.001
	(0.001)	(0.002)	(0.002)
Low piece rate		-0.001	-0.001
× Praise		(0.003)	(0.003)
High piece rate		-0.006	-0.006
× Praise		(0.004)	(0.004)
Low piece rate		0.003	0.003
× Reference point		(0.003)	(0.003)
High piece rate		-0.002	-0.002
× Reference point		(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***	0.017***	0.021***
	(0.001)	(0.002)	(0.003)
N	62026	62026	62026
R^2	0.002	0.002	0.002
R^2 (Within)	0.000	0.000	0.000
R^2 (Between)	0.001	0.003	0.013

Note: The table reports estimation results from random effects panel regressions in which the error rate per fragment and worker is regressed on a set of explanatory variables. "Low piece rate": indicator variable taking the value of one if the treatment used a low piece rate. "High piece rate": indicator variable taking the value of one if the treatment used a high piece rate. "Praise": indicator variable taking the value of one if the treatment praised workers. "Reference point": indicator variable taking the value of one if the treatment set a reference point. "Age": continuous variable measuring a worker's age. "Female": indicator variable taking the value one if the worker is a female. "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": indicator variable taking the value one if the worker uses a mobile device. "Latin": indicator variable variable taking the value of one if the worker has at least some knowledge of Latin. Robust standard errors in parentheses ($* : p < 0.1$, $** : p < 0.05$, $*** : p < 0.01$).

Figure 3: Quantity vs. quality, study 1



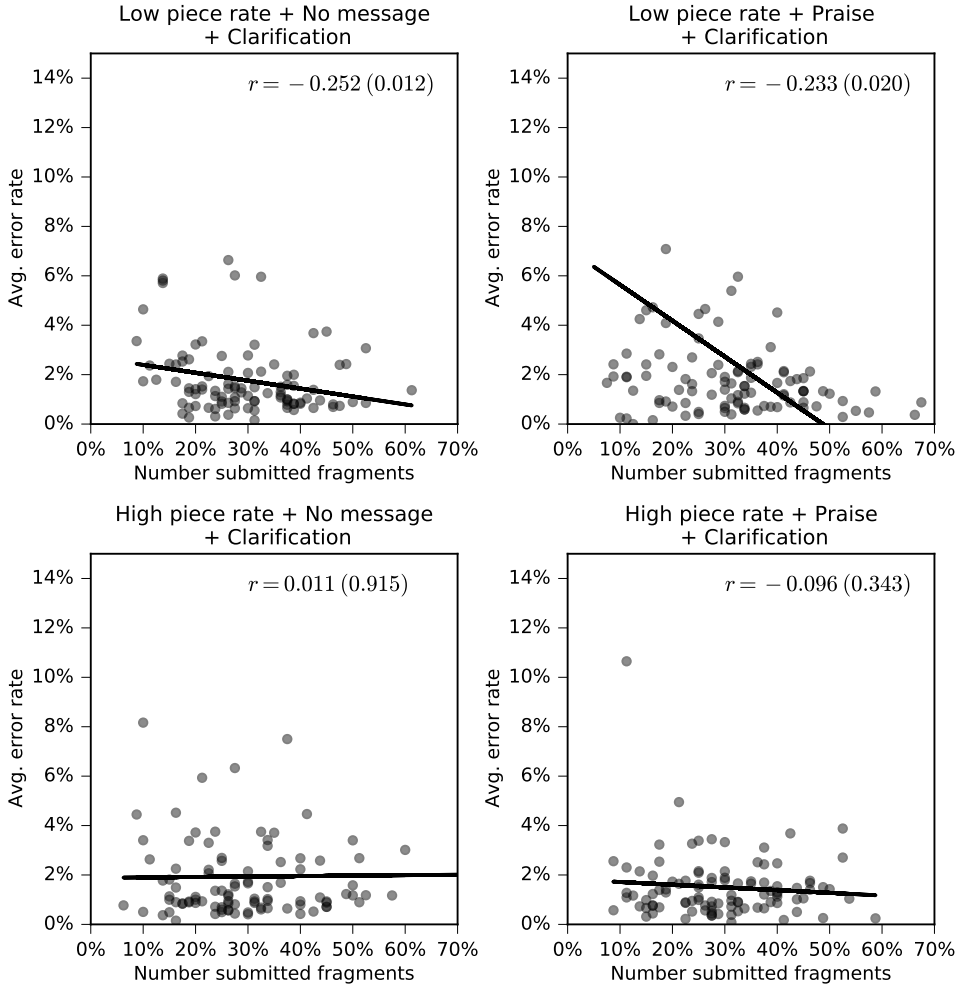
Note: The figure plots the number of submitted fragments per worker as a percentage of the total number of fragments a worker could submit against the average error rate for all submitted fragments per worker for each treatment combination. Indicated as well are the overlaid linear predictions as well as the Pearson correlation coefficient along with p-values (in parentheses).

Table 5: Instrumental variable estimation, study 1

Model	OLS	2SLS	
		1st stage	2nd stage
Dependent variable:	Avg. error rate	Share fragments	Avg. error rate
Share fragments	-0.036***		-0.019
	(0.007)		(0.073)
Age	-0.000	-0.002***	-0.000
	(0.000)	(0.000)	(0.000)
Female	-0.004***	0.009**	-0.004***
	(0.001)	(0.004)	(0.001)
Education	-0.000	0.007***	-0.001
	(0.000)	(0.002)	(0.001)
Mobile device	0.006**	-0.063***	0.007
	(0.002)	(0.011)	(0.005)
Latin	-0.000	0.012**	-0.000
	(0.001)	(0.005)	(0.002)
Constant	0.034***	0.340***	0.029
	(0.004)	(0.011)	(0.025)
Low piece rate		0.025***	
		(0.009)	
High piece rate		0.025***	
		(0.009)	
Praise		-0.016*	
		(0.009)	
Reference point		0.001	
		(0.008)	
Low piece rate		-0.006	
× Praise		(0.013)	
High piece rate		0.001	
× Praise		(0.013)	
Low piece rate		-0.020*	
× Reference point		(0.012)	
High piece rate		-0.025**	
× Reference point		(0.012)	
N	2680	2680	2680
R^2	0.027	0.084	0.024
Partial F-statistic		30.48***	
Wu-Hausman F			0.204
Durbin χ^2			0.204
Sargan χ^2			8.248

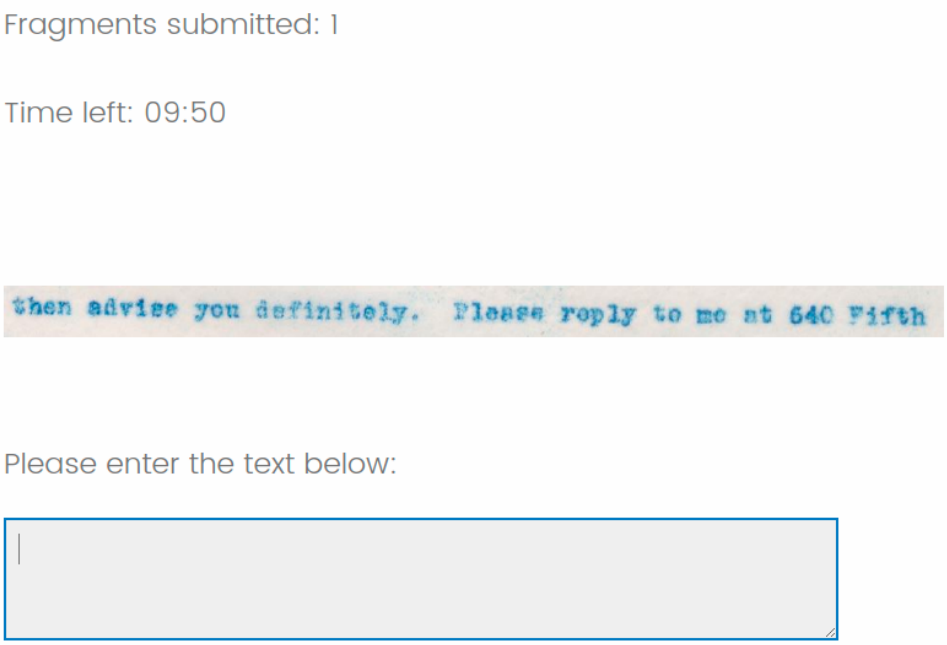
Note: The table reports OLS and 2SLS estimation results for regressions in which the time averaged error rate for each worker is regressed against the number of submitted fragments as a share of the total number of fragments a worker could submit ("Share fragments"). "Age": continuous variable measuring a worker's age. "Female": indicator variable taking the value one if the worker is a female. "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": indicator variable taking the value one if the worker uses a mobile device. "Latin": indicator variable variable taking the value of one if the worker has at least some knowledge of Latin. "Low piece rate": indicator variable taking the value of one if the treatment used a low piece rate. "High piece rate": indicator variable taking the value of one if the treatment used a high piece rate. "Praise": indicator variable taking the value of one if the treatment praised workers. "Reference point": indicator variable taking the value of one if the treatment set a reference point. Robust standard errors in parentheses ($* : p < 0.1$, $** : p < 0.05$, $*** : p < 0.01$).

Figure 4: Quantity vs. quality, clarification treatments only, study 1



Note: The figure plots the number of submitted fragments per worker as a percentage of the total number of fragments a worker could submit against the average error rate for all submitted fragments per worker for each clarification treatment. Indicated as well are the overlaid linear predictions as well as the Pearson correlation coefficients along with p-values (in parentheses).

Figure 5: Screenshot of the work task, study 2



Note: The picture shows an example fragment from the task used in study 2 that workers had to transcribe.

Table 6: Treatment table, study 2

	Non-Goal related CLTs		
Goal related CLTs	No	Yes	All
No	442	438	880
Yes	437	441	878
All	879	879	1758

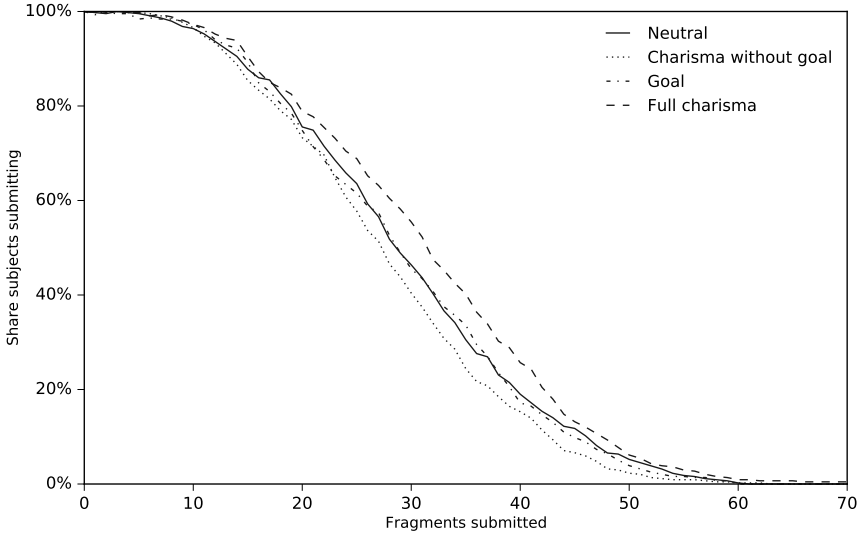
Note: The table gives an overview of the experimental design of study 2 and shows the combination of non-goal related charismatic leadership tactics (CLTs) and goal-related CLT treatment interventions. The number of subjects for each treatment cell is indicated as well.

Table 7: Background characteristics of subjects, study 2

		Age	Education	Female	Mobile device	
Non-goal rel. CLT	**Goal rel. CLT**	**Mean (se)**	**Mean (se)**	**Mean (se)**	**Mean (se)**	**N**
No	No	37.29 (0.54)	4.57 (0.06)	0.51 (0.02)	0.03 (0.01)	442
	Yes	37.78 (0.55)	4.45 (0.06)	0.46 (0.02)	0.06 (0.01)	437
Yes	No	37.74 (0.55)	4.41 (0.06)	0.48 (0.02)	0.03 (0.01)	438
	Yes	37.94 (0.55)	4.52 (0.06)	0.45 (0.02)	0.03 (0.01)	441
All		37.68 (0.27)	4.49 (0.03)	0.47 (0.01)	0.04 (0.00)	1758
Pr(>F)		0.88	0.33	0.27	0.20	

Note: The table reports background characteristics of subjects participating in study 2. 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. Reported are also p-values for the overall regression F-statistic from models in which the respective background characteristic is regressed on all treatment indicator variables.

Figure 6: Fragments submitted, study 2



Note: The figure plots the inverse cumulative distribution function for the number of submitted fragments in all treatments.

Table 8: Treatment effects on quantity, study 2

Model:	I	II	III	IV
Goal	1.720***	-0.310	0.275	0.217
	(0.560)	(0.790)	(0.748)	(0.743)
Charisma	0.274	-1.753**	-1.565**	-1.692**
	(0.560)	(0.765)	(0.734)	(0.734)
Charisma× Goal		4.059***	3.532***	3.626***
		(1.116)	(1.067)	(1.067)
Age			-0.204***	-0.202***
			(0.022)	(0.022)
Female			1.832***	1.860***
			(0.535)	(0.532)
Education			0.503**	0.487**
			(0.205)	(0.204)
Mobile device			-13.456***	-13.253***
			(0.920)	(0.957)
Constant	28.319***	29.328***	34.130***	30.243***
	(0.486)	(0.560)	(1.334)	(1.571)
Group intercepts	No	No	No	Yes
N	1758	1758	1758	1758
R^2	0.005	0.013	0.101	0.125
F	4.758	7.594	45.287	16.560
Pr(>F)	0.009	0.000	0.000	0.000

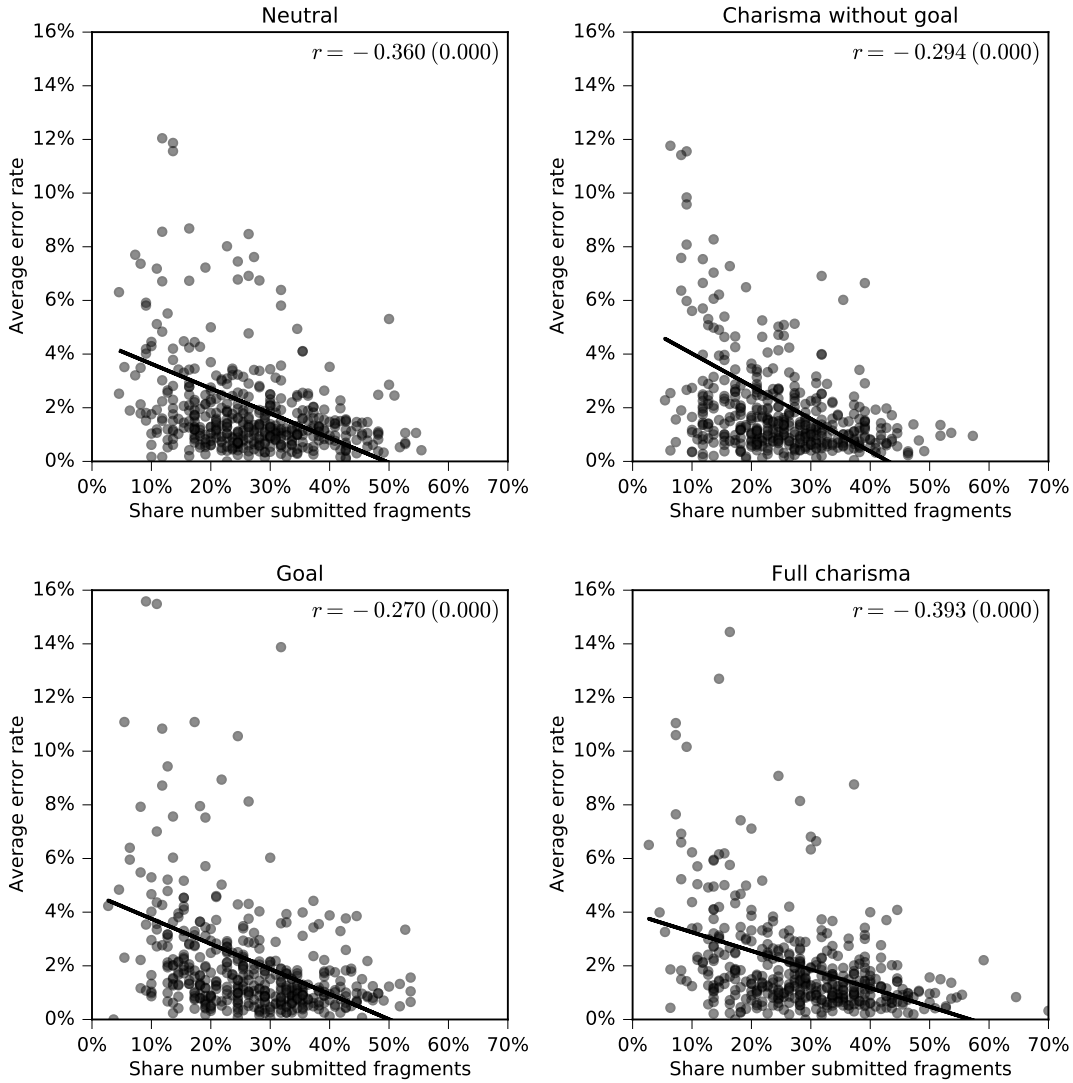
Note: The table reports linear regression results of the number of fragments submitted per worker on a set of explanatory variables. "Goal": indicator variable taking the value of one if the treatment uses goal-related CLTs. "Charisma": indicator variable taking the value of one if the treatment employed non-goal related CLTs. "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 used a mobile device. "Group intercepts": Indicates whether the model specification includes indicator variables for each fragment group (estimate results not reported here). Robust standard errors in parentheses ($* : p < 0.1$, $** : p < 0.05$, $*** : p < 0.01$).

Table 9: Treatment effects on quality, study 2

Model	I	II	III	IV
Goal	-0.000	0.001	0.001	0.001
	(0.002)	(0.003)	(0.002)	(0.002)
Charisma	-0.001	0.001	0.001	0.001
	(0.002)	(0.003)	(0.002)	(0.002)
Charisma×Goal		-0.003	-0.003	-0.003
		(0.003)	(0.003)	(0.003)
Age			-0.000	-0.000
			(0.000)	(0.000)
Female			-0.005***	-0.005***
			(0.002)	(0.002)
Education			0.001	0.001
			(0.001)	(0.001)
Mobile device			0.030***	0.031***
			(0.008)	(0.008)
Constant	0.021***	0.020***	0.019***	0.023***
	(0.002)	(0.002)	(0.004)	(0.005)
Group intercepts	No	No	No	Yes
N	51517	51517	51517	51517
R^2	0.003	0.003	0.004	0.005
R^2 (Within)	0.000	0.000	0.000	0.000
R^2 (Between)	-0.002	0.005	0.046	0.066

Note: The table reports linear regression results of the number of fragments submitted per worker on a set of explanatory variables. "Goal": indicator variable taking the value of one if the treatment uses goal-related CLTs. "Charisma": indicator variable taking the value of one if the treatment employed Non-goal related CLTs. "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 used a mobile device. "Group intercepts": Indicates whether the model specification includes indicator variables for each fragment group (estimate results not reported here). Robust standard errors in parentheses ($* : p < 0.1$, $** : p < 0.05$, $*** : p < 0.01$).

Figure 7: Quality vs. quantity, study 2



Note: The figure plots the number of submitted fragments per worker as a percentage of the total number of fragments a worker could submit against the average error rate 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).

Table 10: Instrumental variable estimation, study 2

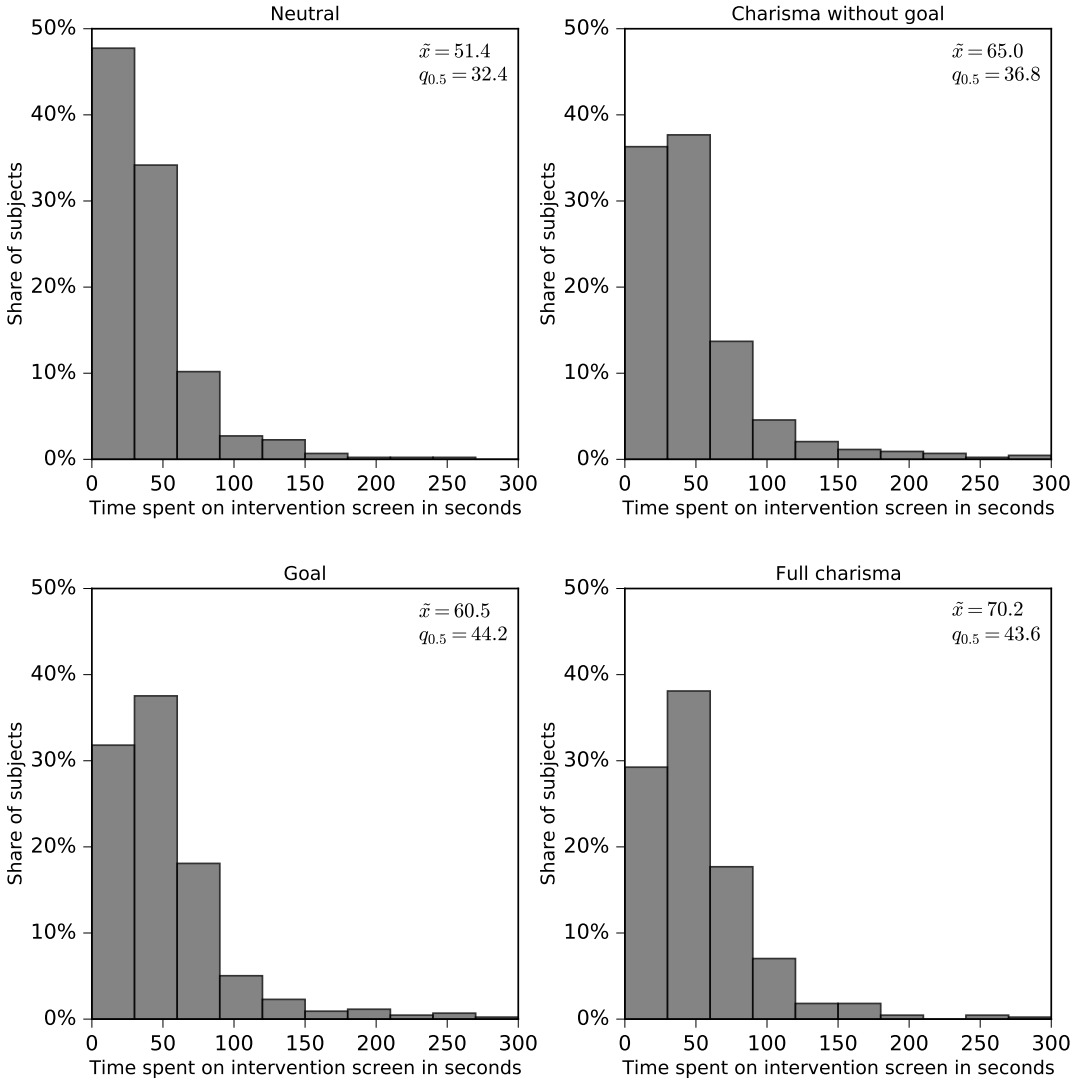
Model	OLS	2SLS	
		1st stage	2nd stage
Dependent variable:	Avg. error rate	Share fragments	Avg. edit ratio
Share fragments	-0.089***		-0.073
	(0.010)		(0.057)
Age	-0.000***	0.002***	-0.000
	(0.000)	(0.000)	(0.000)
Female	-0.004***	0.017***	-0.004***
	(0.002)	(0.005)	(0.002)
Education	0.002***	0.002**	0.002***
	(0.000)	(0.001)	(0.001)
Mobile device	0.020*	-0.122***	0.022*
	(0.011)	(0.008)	(0.012)
Constant	0.047***	0.310***	0.042**
	(0.005)	(0.012)	(0.018)
Charisma		-0.014**	
		(0.007)	
Goal		0.002	
		(0.007)	
Charisma×Goal		0.032***	
		(0.010)	
N	1758	1758	1758
R^2	0.125	0.101	0.122
Partial F-statistic		25.20***	
Wu-Hausman F			0.090
Durbin χ^2			0.090
Sargan χ^2			0.110

Note: The table reports OLS and 2SLS estimation results for regressions in which the time averaged error rate for each worker is regressed against the number of submitted fragments as a share of the total number of fragments a worker could submit ("Share fragments"). "Age": continuous variable measuring a worker's age. "Female": indicator variable taking the value one if the worker is a female. "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": indicator variable taking the value one if the worker uses a mobile device. "Goal": indicator variable taking the value of one if the treatment uses goal-related CLTs. "Charisma": indicator variable taking the value of one if the treatment employed non-goal related CLTs. Robust standard errors in parentheses ($* : p < 0.1$, $** : p < 0.05$, $*** : p < 0.01$).

Appendix

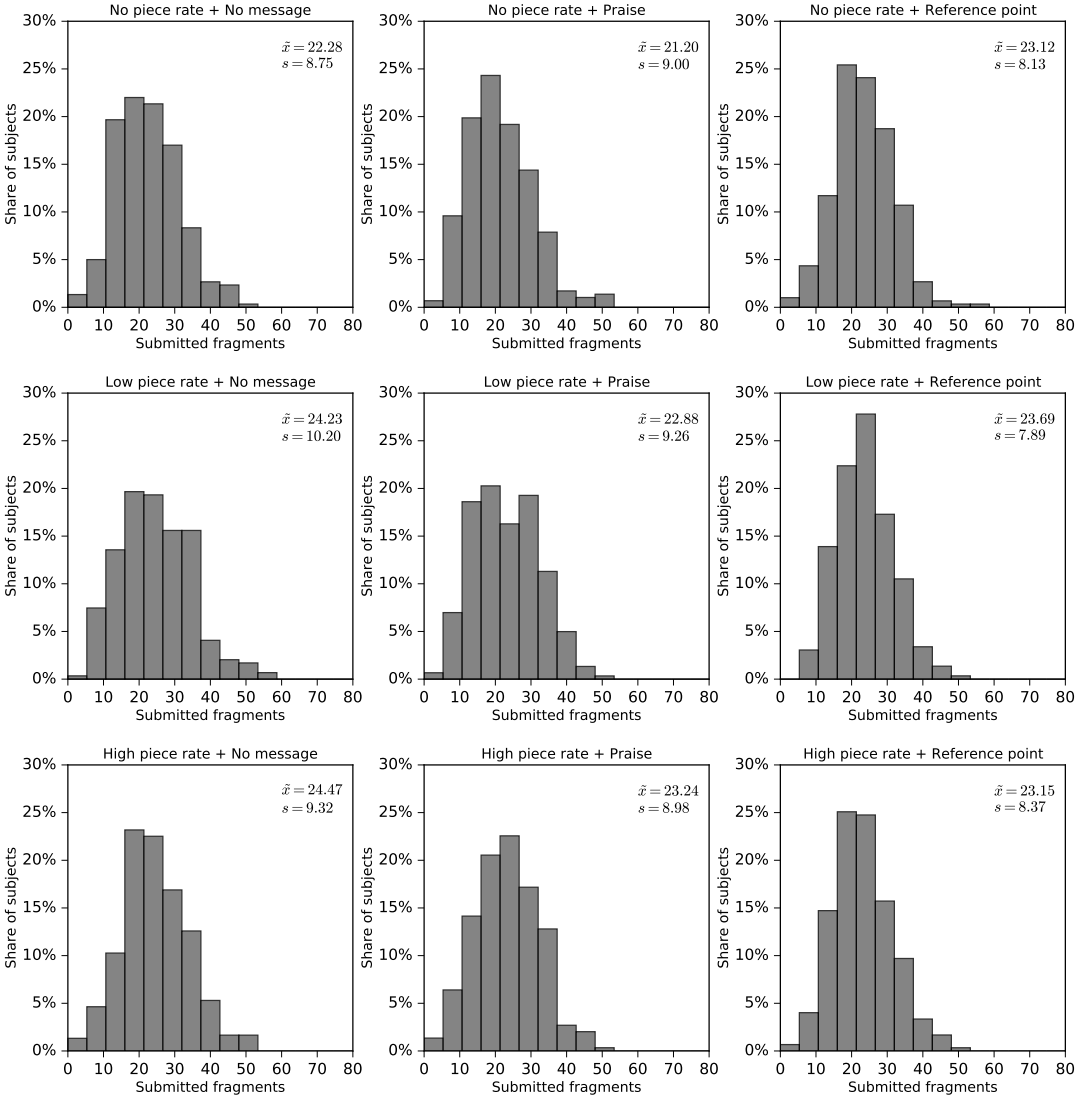
6.1 Additional tables and figures

Figure S1: Time spent on intervention screen, study 1



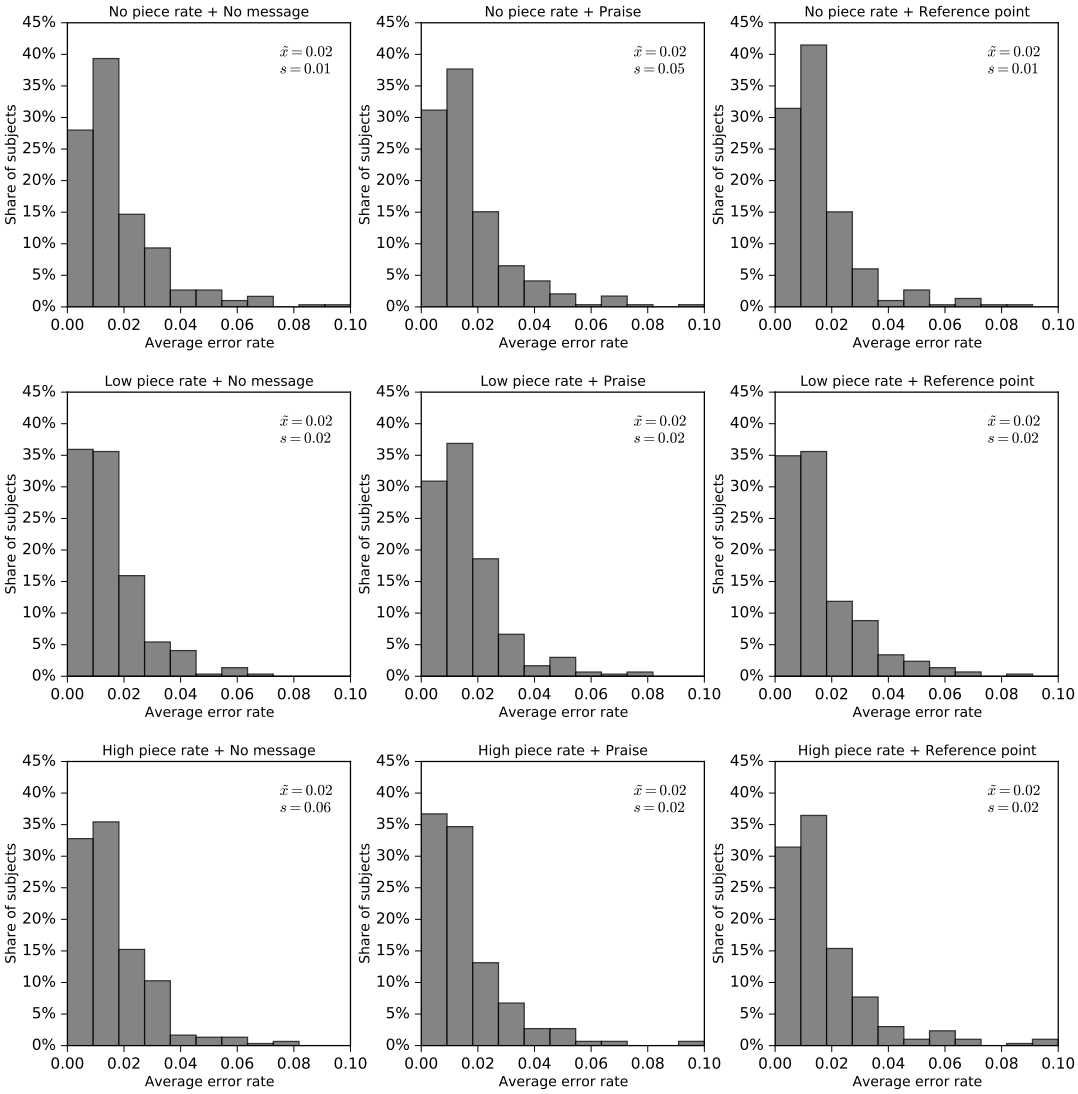
Note: The figure shows the histogram of time spent on the intervention screen for the treatments with a praise message (left panel) and the treatments with a reference point message (right panel). The mean ($\bar{x}$) and median ($q_{0.5}$) time spent on the intervention screen are reported in each panel as well.

Figure S2: Histogram fragments submitted, study 1



Note: The figure shows the histogram for the number of submitted fragments in all treatments in study 1. Indicated as well are the mean ($\bar{x}$) and standard deviation (s).

Figure S3: Histogram error rate, study 1



Note: The figure shows the histogram for the average error rate per worker in all treatments in study 1. Indicated as well are the mean ($\bar{x}$) and standard deviation (s).

Table S1: Quality vs. quantity, study 1

Model	I	II	III	IV
Share fragments	-0.036***	-0.036***	-0.030***	-0.089*
	(0.007)	(0.007)	(0.006)	(0.048)
Constant	0.028***	0.032***	0.028***	0.049***
	(0.003)	(0.005)	(0.003)	(0.018)
Intercepts	No	Yes	No	Yes
Slopes	No	No	Yes	Yes
N	2680	2680	2680	2680
R^2	0.018	0.021	0.019	0.029
F	24.912	5.013	4.062	4.289
Pr(>F)	0.000	0.000	0.000	0.000

Note: The table reports estimation results for regressions in which the time averaged error rate per worker is regressed against the number of submitted fragments per worker as a percentage of the total number of fragments a worker could submit ("Share fragments"). Indicated as well is whether the model specification includes indicator variables for each treatment ("Intercepts") and treatment specific slopes ("Slopes") (estimate results not reported here). Robust standard error in parentheses ($* : p < 0.1$, $** : p < 0.05$, $*** : p < 0.01$).

Table S2: Quality vs. quantity, clarification treatments, study 1

Model	I	II	III	IV
Share fragments	-0.036***	-0.036***	-0.030***	-0.089*
	(0.007)	(0.007)	(0.006)	(0.048)
Constant	0.028***	0.032***	0.028***	0.049***
	(0.003)	(0.005)	(0.003)	(0.018)
Intercepts	No	Yes	No	Yes
Slopes	No	No	Yes	Yes
N	2680	2680	2680	2680
R^2	0.018	0.021	0.019	0.029
F	24.912	5.013	4.062	4.289
Pr(>F)	0.000	0.000	0.000	0.000

Note: The table reports estimation results for regressions in which the time averaged error rate per worker is regressed against the number of submitted fragments per worker as a percentage of the total number of fragments a worker could submit ("Share fragments"). Indicated as well is whether the model specification includes indicator variables for each treatment ("Intercepts") and treatment specific slopes ("Slopes") (estimate results not reported here). Robust standard error in parentheses ($* : p < 0.1$, $** : p < 0.05$, $*** : p < 0.01$).

Table S3: Quality vs. quantity, clarification vs. no clarification, study 1

Model	I	II	III
Share fragments	-0.042***	-0.040***	-0.041***
	(0.013)	(0.012)	(0.011)
Clarification		0.005	0.004
		(0.013)	(0.013)
Share fragments × Clarification		-0.010	-0.007
		(0.037)	(0.037)
Constant	0.031***	0.030***	0.038***
	(0.005)	(0.004)	(0.005)
Controls	No	No	Yes
N	1591	1591	1591
R2	0.018	0.019	0.026
F	11.169	4.927	4.730
Pr(>F)	0.001	0.002	0.000

Note: The table reports linear regression results for regressing the averaged error rate for all fragments on the number of submitted fragments per worker as a percentage of the total number of fragments a worker could submit ("Share fragments"). "Clarification": indicator variable taking the value one if workers received the information that we would not check the quality of their submitted fragments. 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, study 1

Model	I	II	III	IV	V
High piece rate	0.094			0.240	0.005
	(0.479)			(0.801)	(0.770)
Praise		-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 × Praise				0.119	0.550
				(1.095)	(1.062)
High piece rate × Clarification				-0.087	0.469
				(1.600)	(1.528)
Praise × Clarification				2.473	2.732*
				(1.601)	(1.557)
High piece rate × Praise × 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

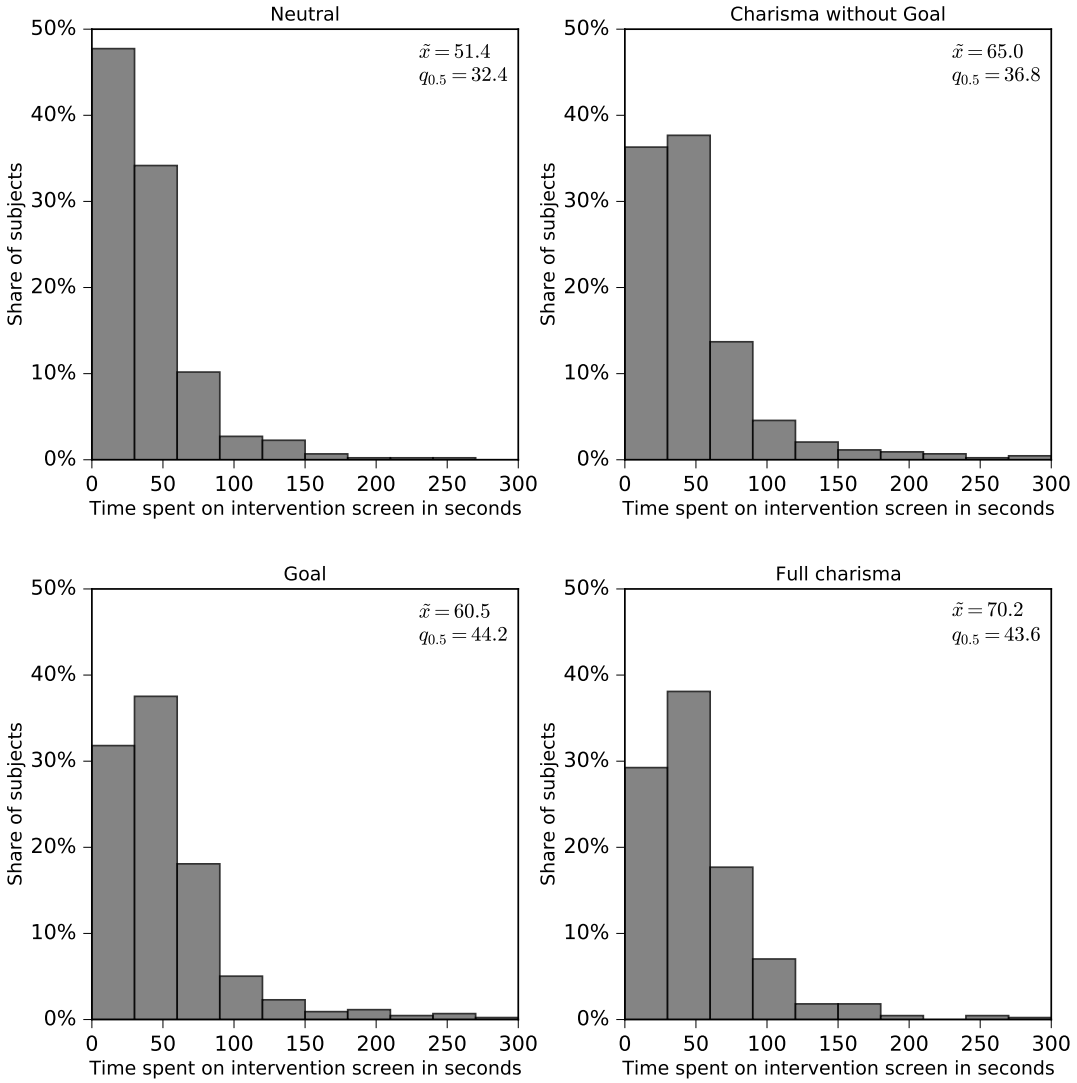
Note: The table reports linear regression estimation results from regressing the the number of fragments submitted per worker on a set of explanatory variables. "High piece rate": indicator variable taking the value one if workers received a high piece rate. "Praise": indicator variable taking the value one if workers received praise prior to work. "Clarification": indicator variable taking the value one if workers received the information that we would not check the quality of their submitted fragments. 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, study 1

Model	I	II	III	IV	V
High piece rate	0.001			0.005	0.005
	(0.002)			(0.003)	(0.003)
Praise		-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 × Praise				-0.005	-0.005
				(0.005)	(0.005)
High piece rate × Clarification				-0.003	-0.004
				(0.005)	(0.005)
Praise × Clarification				0.007	0.006
				(0.007)	(0.007)
High piece rate × Praise × 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 estimation results from regressing the error rate per fragment on a set of explanatory variables. "High piece rate": indicator variable taking the value one if workers received a high piece rate. "Praise": indicator variable taking the value one if workers received praise prior to work. "Clarification": indicator variable taking the value one if workers received the information that we would not check the quality of their submitted fragments. Controls include variables for workers' age, gender, education, use of mobile device and knowledge of Latin. Standard errors in parentheses ($* : p < 0.1$, $** : p < 0.05$, $*** : p < 0.01$).

Figure S4: Time spent on intervention screen, study 2



Note: The figure shows the histogram of time spent on the intervention in all treatments for study 2. The mean ($\bar{x}$) and median ($q_{0.5}$) time spent on the intervention screen are reported in each panel as well.

Table S6: Categories and tactics, study 2

Framing and vision	C1	metaphors or similis, to simplify the message, stir emotions, and make parallels between symbolic meanings and realities more salient
	C2	rhetorical questions, to create intrigue and suspense, and direct attention to seeking the answer
	C3	stories and anecdotes, to simplify the message, trigger imagery and recall, engender identification with characters in the story, and identify a relevant moral
	C4	contrasts, to define what should be done versus what should not be done by showcasing the right way versus a wrong way
	C5	three-part lists, to provide sufficient proof or completeness
Substance	C6	expressing moral conviction, to focus attention on moral justification and on doing what is morally right
	C7	expressing the sentiments of the collective, to engender identification (via similarity) with the leader
	C8	setting high and ambitious goals, to make followers feel competent and focus their effort on a target
	C9	creating confidence that goals can be achieved, to raise follower confidence and make them more likely to exert effort

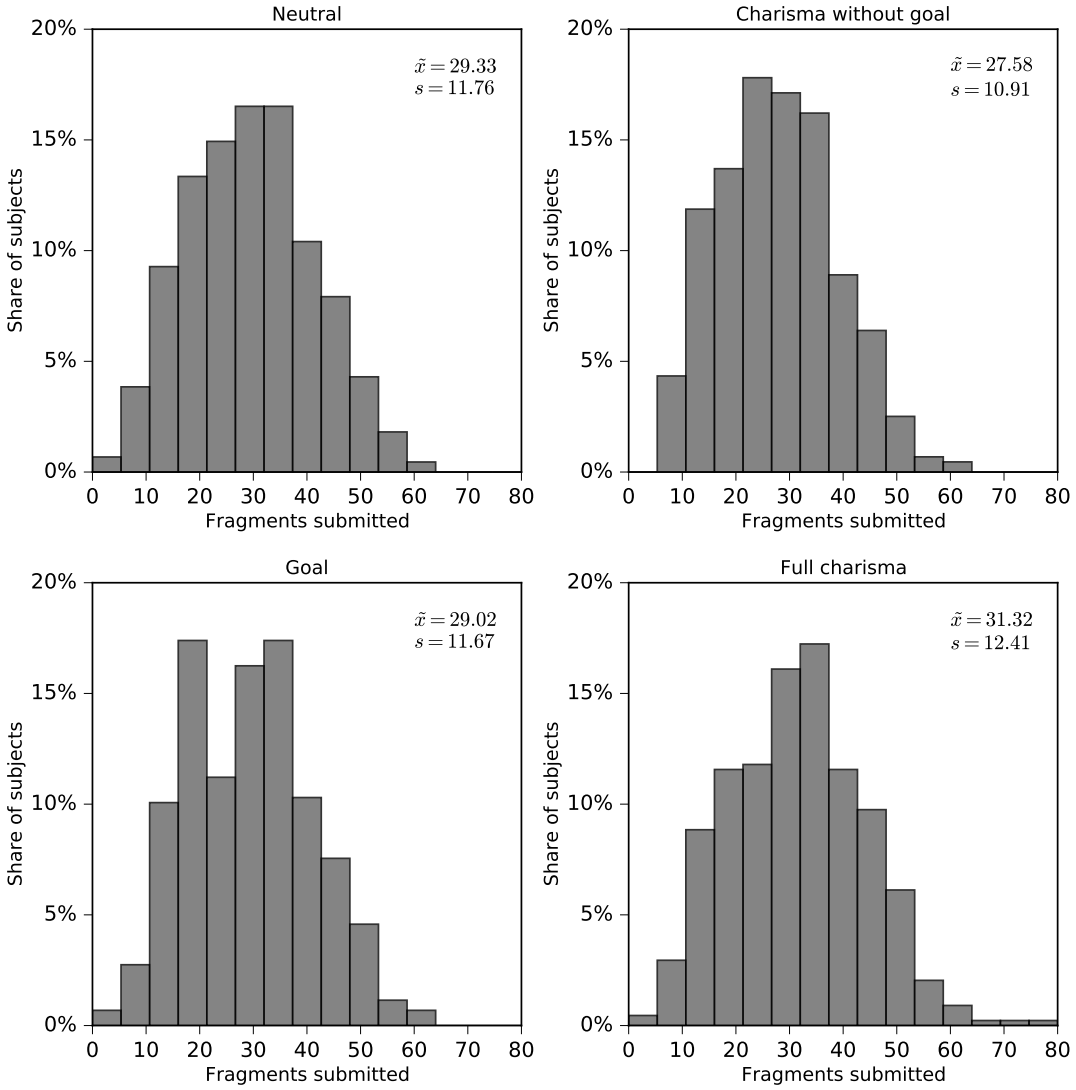
Table S7: Coding of sentences for *Neutral* and *Goal*, study 2

No.	Treatment	Sentence	C1	C2	C3	C4	C5	C6	C7	C8	C9
1	Neutral & Goal	Welcome to this HIT.									
2	Neutral & Goal	Your job will be to transcribe text from historic documents from the Frick Collection and Frick Art Reference Library Archives.									
3	Neutral & Goal	You will see fragments of these documents on the screen and we kindly ask you to type the text into the blank space below the fragment using your keyboard.									
4	Neutral & Goal	We care about quantity and quality of work.									
5	Neutral & Goal	You will be paid no matter how many fragments you submit.									
6	Neutral & Goal	The transcriptions you are going to create will become searchable data points in a large database.									
7	Neutral & Goal	Your effort will help the project.						1			
8	Neutral & Goal	Each fragment you manage to transcribe will translate into one more data point.									
9	Neutral & Goal	Together with hundreds of other MTurkers working on this HIT, your work will contribute to preserve and build knowledge of past events.							1		
10	Neutral & Goal	This data can then be accessed by scholars, students or the public in general for study purposes.									
11	Neutral & Goal	We ask you to work hard and diligently as well as to produce high quality output.					1				
12	Goal	In similar HITs, MTurkers submitted roughly 25 fragments on average.								1	1
13	Goal	We ask you to aim for at least 34 fragments.								1	
14	Goal	This is a challenging goal but because you have already worked on many HITs and earned an excellent approval rate, we are confident that you will be able to meet or even exceed this goal.									1
15	Neutral & Goal	Below, you see an example of the task.									
16	Neutral & Goal	In the actual assignment, after you have submitted the text, a new fragment will appear on your screen.									
17	Neutral & Goal	In total, you will have to work on the assignment for 10 minutes.									
18	Neutral & Goal	After finishing the assignment, you will be taken to a short questionnaire.									

Table S8: Coding of sentences for *Charisma without goal* and *Full charisma*, study 2

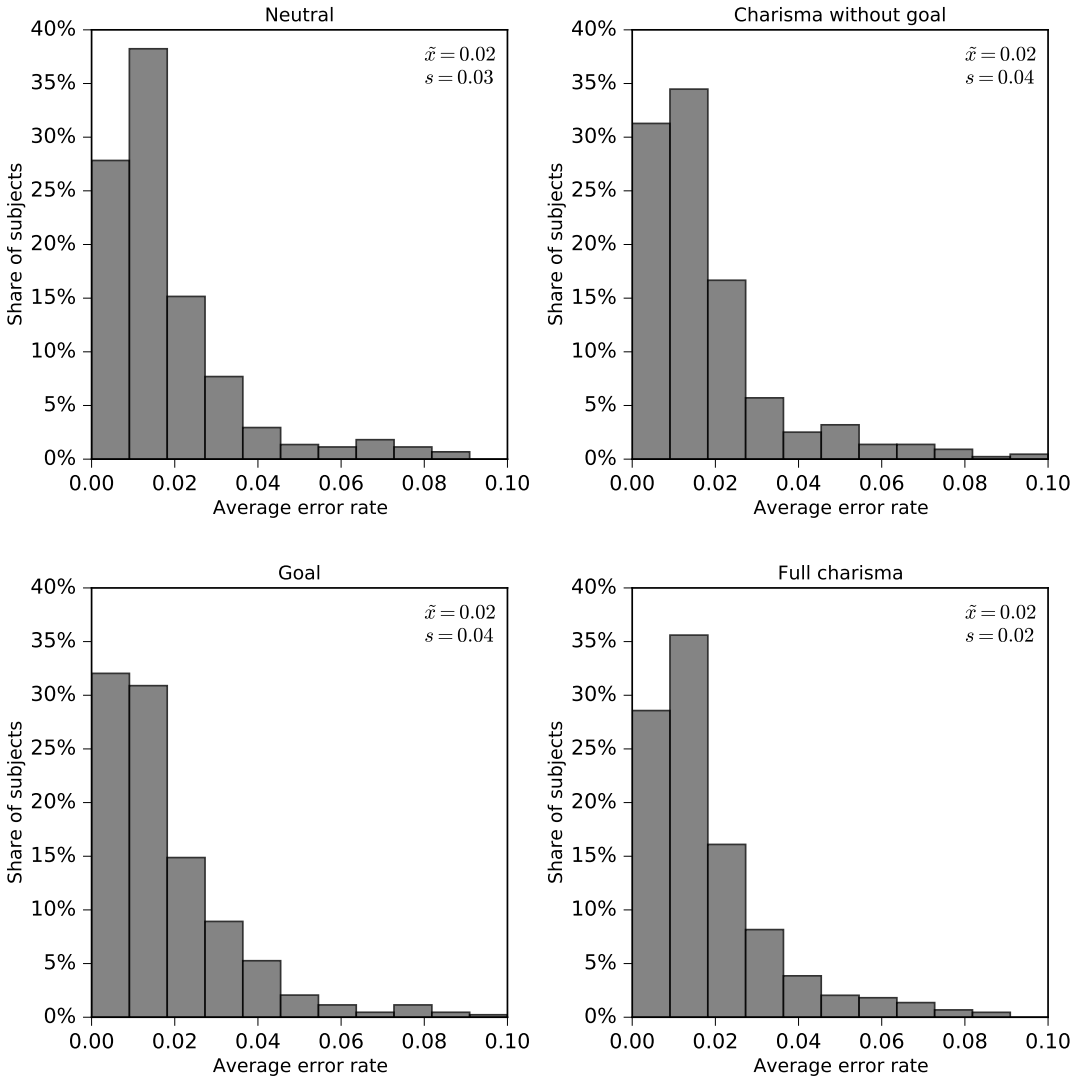
No.	Treatment	Sentence	C1	C2	C3	C4	C5	C6	C7	C8	C9
1	Char. w/o Goal & Full Char.	Welcome to this HIT.									
2	Char. w/o Goal & Full Char.	Your job will be to transcribe text from historic documents from the Frick Collection and Frick Art Reference Library Archives.									
3	Char. w/o Goal & Full Char.	You will see fragments of these documents on the screen and we kindly ask you to type the text into the blank space below the fragment using your keyboard.									
4	Char. w/o Goal & Full Char.	We care about quantity and quality of work.									
5	Char. w/o Goal & Full Char.	You will be paid no matter how many fragments you submit.									
6	Char. w/o Goal & Full Char.	The transcriptions you create will become searchable data facilitating learning and research around the world.						1			
7	Char. w/o Goal & Full Char.	You might think, will my extra effort really help?		1							
8	Char. w/o Goal & Full Char.	Yes, it will!			1			1			
9	Char. w/o Goal & Full Char.	Each fragment is like a little piece of a puzzle; together with hundreds of other MTurkers, you will put the puzzle together.	1						1		
10	Char. w/o Goal & Full Char.	You can bring history to life and keep it alive.	1								
11	Char. w/o Goal & Full Char.	Just like historians, you contribute to preserve and build the public knowledge of past events.			1						
12	Char. w/o Goal & Full Char.	So, we ask you to jump in and work hard, work diligently, and produce high-quality output.					1				
13	Char. w/o Goal & Full Char.	Not only do you benefit from this job; so too will students, scholars, and the public at large.						1	1		
14	Full Char.	In similar HITs, MTurkers submitted roughly 25 fragments on average.								1	1
15	Full Char.	We ask you to aim for at least 34 fragments.								1	
16	Full Char.	This is a challenging goal but because you have already worked on many HITs and earned an excellent approval rate, we are confident that you will be able to meet or even exceed this goal.									1
17	Char. w/o Goal & Full Char.	Below, you see an example of the task.									
18	Char. w/o Goal & Full Char.	In the actual assignment, after you have submitted the text, a new fragment will appear on your screen.									
19	Char. w/o Goal & Full Char.	In total, you will have to work on the assignment for 10 minutes.									
20	Char. w/o Goal & Full Char.	After finishing the assignment, you will be taken to a short questionnaire.									

Figure S5: Histogram fragments submitted, study 2



Note: The figure shows the histogram for the number of submitted fragments in all treatments in study 2. Indicated as well are the mean ($\bar{x}$) and standard deviation (s).

Figure S6: Histogram error rate, study 2



Note: The figure shows the histogram for the average error rate per worker in all treatments in study 2. Indicated as well are the mean ($\bar{x}$) and standard deviation (s).

Table S9: Quality vs. quantity, study 2

Model	I	II	III	IV
Share fragments	-0.092***	-0.092***	-0.094***	-0.092***
	(0.010)	(0.010)	(0.011)	(0.017)
Constant	0.046***	0.046***	0.046***	0.046***
	(0.003)	(0.003)	(0.003)	(0.006)
Intercepts	No	Yes	No	Yes
Slopes	No	No	Yes	Yes
N	1758	1758	1758	1758
R2	0.093	0.094	0.095	0.097
F	85.364	28.483	23.595	17.600
Pr(>F)	0.000	0.000	0.000	0.000

Note: The table reports estimation results for regressions in which the time averaged error rate per worker is regressed against the number of submitted fragments per worker as a percentage of the total number of fragments a worker could submit ("Share fragments"). Indicated as well is whether the model specification includes indicator variables for each treatment ("Intercepts") and treatment specific slopes ("Slopes") (estimate results not reported here). Robust standard error in parentheses ($* : p < 0.1$, $** : p < 0.05$, $*** : p < 0.01$).

6.2 Instructions study 1

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**.][21]

[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**.][22]

{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 Praise 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!

{New page for Reference point treatments}

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

[21] These treatments where pooled with the main treatments, compare footnote 2.
[22] We discuss these treatments in Section 3.3.3.

6.3 Instructions study 2

You will be paid a **fixed compensation of $3** for working on this project. The compensation will be sent to you within two days after the completion of this HIT.

{New page}

Neutral treatment
Welcome to this HIT. Your job will be to transcribe text from historic documents from the Frick Collection and Frick Art Reference Library Archives. You will see fragments of these documents on the screen and we kindly ask you to type the text into the blank space below the fragment using your keyboard. We care about quantity and quality of work. You will be paid no matter how many fragments you submit.

The transcriptions you are going to create will become searchable data points in a large database. Your effort will help the project. Each fragment you manage to transcribe will translate into one more data point. Together with hundreds of other MTurkers working on this HIT, your work will contribute to preserve and build knowledge of past events. This data can then be accessed by scholars, students or the public in general for study purposes. We ask you to work hard and diligently as well as to produce high quality output.

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

Goal Treatment (same as *Neutral* plus paragraph for goals)
Welcome to this HIT. Your job will be to transcribe text from historic documents from the Frick Collection and Frick Art Reference Library Archives. You will see fragments of these documents on the screen and we kindly ask you to type the text into the blank space below the fragment using your keyboard. We care about quantity and quality of work. You will be paid no matter how many fragments you submit.

The transcriptions you are going to create will become searchable data points in a large database. Your effort will help the project. Each fragment you manage to transcribe will translate into one more data point. Together with hundreds of other MTurkers working on this HIT, your work will contribute to preserve and build knowledge of past events. This data can then be accessed by scholars,

students or the public in general for study purposes. We ask you to work hard and diligently as well as to produce high quality output.

In similar HITs, MTurkers submitted roughly 25 fragments on average. We ask you to aim for at least 34 fragments. This is a challenging goal but because you have already worked on many HITs and earned an excellent approval rate, we are confident that you will be able to meet or even exceed this goal.

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

Charisma without goal treatment
Welcome to this HIT. Your job will be to transcribe text from historic documents from the Frick Collection and Frick Art Reference Library Archives. You will see fragments of these documents on the screen and we kindly ask you to type the text into the blank space below the fragment using your keyboard. We care about quantity and quality of work. You will be paid no matter how many fragments you submit.

The transcriptions you create will become searchable data facilitating learning and research around the world. You might think, will my extra effort really help? Yes, it will! Each fragment is like a little piece of a puzzle; together with hundreds of other MTurkers, you will put the puzzle together. You can bring history to life and keep it alive. Just like historians, you contribute to preserve and build the public knowledge of past events. So, we ask you to jump in and work hard, work diligently, and produce high-quality output. Not only do you benefit from this job; so too will students, scholars, and the public at large.

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

Full charisma treatment (same as *Charisma without goal* plus paragraph for goals from *Goal* treatment)
Welcome to this HIT. Your job will be to transcribe text from historic documents from the Frick Collection and Frick Art Reference Library Archives. You will see fragments of these documents on the screen and we kindly ask you to type the text into the blank space below the fragment using your keyboard. We care about quantity and quality of work. You will be paid no matter how many

fragments you submit.

The transcriptions you create will become searchable data facilitating learning and research around the world. You might think, will my extra effort really help? Yes, it will! Each fragment is like a little piece of a puzzle; together with hundreds of other MTurkers, you will put the puzzle together. You can bring history to life and keep it alive. Just like historians, you contribute to preserve and build the public knowledge of past events. So, we ask you to jump in and work hard, work diligently, and produce high-quality output. Not only do you benefit from this job; so too will students, scholars, and the public at large.

In similar HITs, MTurkers submitted roughly 25 fragments on average. We ask you to aim for at least 34 fragments. This is a challenging goal but because you have already worked on many HITs and earned an excellent approval rate, we are confident that you will be able to meet or even exceed this goal.

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