Complex tasks and motivating crowdworkers – an experiment

Maja Adena (WZB) and Julian Harke (WZB)

14 January 2021[1]

Abstract:

Crowdsourcing has become more popular in recent years. However, not much is known about how to motivate crowdworkers except for the use of monetary incentives. This becomes even more important when they are conducting tasks that require high engagement but for which the objective quality is difficult to assess and thus performance-based payments difficult to implement. Such tasks might include video, picture, or text classifications and ratings. Per piece payments may increase the speed but actually reduce the quality. In an online experiment, participants have to rate articles. We study the use of (i) recognition and appreciation which, due to the nature of the task, cannot be conditioned on the relevant performance; (ii) positive versus negative framing regarding the payment—depending on the number of attention checks answered correctly the baseline payment gets increased or reduced. We use different more and less objective performance measures as well as measures of objective and subjective crowdworkers' motivation and subjective job satisfaction to assess the effectiveness of our treatments.

Task:

Participants perform a task in which they have to read a newspaper article about a natural disaster and answer an array of questions about it. We ask readers about their subjective perceptions of the articles (emotions). It also asks whether specific information is included, for example, individual stories or statistics about deaths and injuries. If the latter questions are answered in the affirmative, the participants are expected to select the respective parts of the articles. In addition, the participants are asked to select all words relating to the natural disaster in the article.

There are three attention checks that the participants have to answer correctly in order to be paid an additional bonus of *£0.75*. All attention checks are easy to answer and have an objectively true answer. On the first three of four pages with the questions about the article, there is one attention check positioned between the questions. Each time the participant fails an attention check, their bonus payment gets reduced by *£0.25.* At this point participants receive a notification and the mistake is explained (question, the answer given by the participant, and the correct answer). More specifically, the first and third attention check require the participant to move the slider until the question text turns green. The second attention check requires the participant to select "yes" and subsequently fill in the word "cycling". The task concludes with an array of questions about subjective perceptions regarding the task that pertain to job satisfaction and worker motivation. The total payment depends on the length of the article and the number of correctly answered attention checks.

Treatments:

We follow a 3x2 between-subjects design. Each article will be read and rated by 6 participants such that all six treatments are implemented within the same article.

Treatment differences in the first dimension concern the inclusion of recognition: Control treatment (C) is the treatment without any specific recognition; treatment R includes four phrases that express recognition of the work done; and treatment A includes four appreciation phrases.

[1] No data collection has been initialized prior to this date.

After finishing each of the four pages with the questions about the article participants see a screen with a short phrase for 1.7 second.[2] The control group, however, sees a blank page for the same amount of time instead. If participants did not answer the respective attention check correctly, they see the notification that they failed the attention check instead of the screen with a phrase.

Note that even if the recognition/appreciation is conditional on answering the attention checks correctly (except the last instance), they are not related to the relevant performance regarding the actual task—article rating. Therefore, we speak about performance unrelated recognition.

C Control

[Blank pages] x 4

R recognition

1. *Great work!*
2. *You did a good job!*
3. *Nice job!*
4. *Well done!*

A Appreciation

1. *Thank you!*
2. *Your help makes a difference!*
3. *We appreciate your support!*
4. *Your work matters!*

The second dimension consist of positive versus negative framing regarding the performance-dependent part of the payment. The participants receive this information upfront and can look it up at each stage of the experiment ("Payment" button in the lower right corner of each page).

Treatment L (loss domain)

Payment

*If you finish the task you will be paid up to **£1.65/1.95/2.25**.[3] Your final payment depends on you meeting the quality requirements. The quality requirements are based on three simple questions that have a clear and objectively correct answer. These questions are positioned between other questions and tasks. Each time you fail to answer the quality check question correctly, your payment will be reduced by **£0.25**. The maximum reduction equals thus to **£0.75**, in which case you will be only paid **£0.90/1.20/1.50**.*

After the first [second] {third} failed attention check:

You did not answer the quality check question correctly. Your final payment of £1.65/1.95/2.25 [1.40/1.70/2.00] {1.15/1.45/1.75} will be reduced by £0.25.

Treatment G (gain domain)

Payment

*If you finish the task you will be paid at least **£0.90/1.20/1.50**. Your final payment depends on you meeting the quality requirements. The quality requirements are based on three simple questions that have a clear and objectively correct answer. Those questions are positioned*

[2] An average reader needs approximately 0.3 seconds for one word, that is 1.5 seconds for a phrase with 5 words. We increase the time a little bit in order to allow slower readers still to understand the phrase. Due to preselection of participants with or in higher education, we think that this should be sufficient.

[3] We divided the articles in 3 groups depending on length. The different baseline amounts depend on the length of the article: short/middle/long.

between other questions and tasks. Each time you correctly answer the quality check question, your payment will be raised by ***£0.25****. The maximum additional payment equals thus to* ***£0.75*** *in which case you will be paid* ***£1.65/1.95/2.25****.*

After the first {second}[third] failed attention check:

You did not answer the quality check question correctly. Your final payment of £0.90/1.20/1.50 {1.15/1.45/1.75}[1.40/1.70/2.00] will not be raised by £0.25.

Main hypotheses:
H1 a-f: recognition/appreciation increases worker motivation / job satisfaction / performance

Additional hypotheses:
H2: [Overconfidence and inattention] Better performance in A than in R.

> Explanation: Potential opposite effects of overconfidence resulting from too much praise for work done will lead to inattention.

H3: [work to rule] The probability of mistakes is lower in treatment L than in G for objectively verifiable tasks and higher for less objective ones

> Explanation: In the loss domain, participants are more likely to "work to rule" - tasks that can be objectively validated will be correctly completed but less effort will be put in creative or tasks that cannot be objectively assessed.

H4: [Experience] More experienced participants react less to recognition/appreciation (we use a median split).
H5: There is no interaction effect.

> Explanation: We have no prior expectation about the interaction effect.

Outcomes:

1. [job satisfaction and worker motivation / subjective measures] <u>scores based on responses to following questions asked after the experiment</u>:
   a. [satisfaction] Now, we would like to ask you some additional questions. Did you find the task (responses given using a slider and recorded on a scale from 0 ("not at all") to 100 ("very much"))
      i. Interesting?
      ii. Challenging?
      iii. Fun?
      iv. Boring? (reverse)
      v. Inspiring?
   b. [motivation] Why did you put effort into this task? (adapted from Multidimensional Work Motivation Scale, Gagné et al. (2015), responses given using a slider and recorded on a 101-points scale from "not at all" to "very much"))
      - [Am2] I put little effort into this task because I didn't think that it was worth it. (reverse)
      - [Ext-Mat1] I put effort into this task because others will only reward me financially if I put enough effort into this task. (reverse)
      - [Introj4] I put effort into this task because otherwise I would have felt bad about myself.
      - [Ident2] Putting effort into this task aligns with my personal values.

    c. [satisfaction] Are you satisfied with the payment scheme for the task? (responses given using a slider and recorded on a 101-points scale from “not at all” to “very much”))

2. [worker motivation / objective measure] number of words/phrases selected within one question. In the task, the participants are asked whether specific information is contained in the article. Example: “Does the article report statistics about the number of people killed by the natural disaster?” If the participant chooses yes, they are asked to select the respective part of the article. They may select several parts. Similarly, the participants are asked to select all words pertaining to natural disaster in the article. We do not ask for a specific number of words/sentences. It is possible (and potentially sufficient) to select one (or zero when choosing “No”) and continue with the next question. We believe that participants who select more words/sentences per question show higher motivation. Note that more words selected do not necessarily mean that they are correct.
3. [worker motivation and performance / objective measure] consistency. We believe that motivated workers will pay better attention to the task and thus will show more consistency in their answers. For that reason, we will use
    a. Questions about emotions: The answers should be more similar within the set of positive emotions and within the set of negative emotions than between the two sets.
    b. Questions about job satisfaction: The answer to the question “fun” should be close to the opposite of “boring”
4. [performance / objective measure] probability of a mistake in our objective attention checks number 2 and 3 (i.e. only after the first and second recognition/appreciation phrase was shown).[4] For the recognition treatments, we will use later attention checks as an objective measure of performance.
5. [performance / objective measure] probability of a mistake in all objective attention checks. For the hypothesis about the loss/gain domain we will use all attention checks as objective performance measure.
6. [performance / less objective measure] yes/no questions with text highlighting which accuracy is less obvious but mostly verifiable. For example, “Does the text contains references to an authority?” If the participant chooses yes, then she has to select the respective text. Since we cannot check all responses, we will use the following rule of thumb: if the majority of the participants gives the same answer to a question, we assume that the responses are correct. If one, two or three participants choose “yes” as an answer and select a corresponding part of the text, we will do some manual checks (depending on the number, potentially a subsample only) or exclude those ambiguous cases from the analysis. This measure will be used for the hypothesis about the loss/gain domain.

The measures number 1, 3, 4, and 6 will be used for interaction effects.

Sample:
around 4,500 crowdworkers on Prolific. We will apply the following pre-screening criteria and expect the following final sample size of potential participants:

[4] We will correct for the cases in which participants make a mistake in the first attention check.

Variable	description	criterium	Available sample (participants active in the past 90 days) (19.11.2020)
Geographic variables/ Current country of residence		UK	49,819
Languages/first language		English	39,132
Education/Highest education completed		Undergraduate degree (BA/BSc/other); Graduate degree (MA/MSc/MPhil/other); Doctorate degree (PhD/other)	18,965
Basic demographic variables/Age		18-100= all categories	18,915
Basic demographic variables/sex		male or female	18,834
Socioeconomic variables/ Socioeconomic Status	Participants were asked the following question: Think of a ladder (see image) as representing where people stand in society. At the top of the ladder are the people who are best off - those who have the most money, most education and the best jobs. At the bottom are the people who are worst off - who have the least money, least education and the worst jobs or no job. The higher up you are on this ladder, the closer you are to people at the very top and the lower you are, the closer you are to the bottom. Where would you put yourself on the ladder? Choose the number whose position best represents where you would be on this ladder.	1-10, all categories	18,547
Socioeconomic variables/ Household Size		all categories	18,467
Socioeconomic variables/ Household Income (GBP)	What is your total household income per year, including all earners in your household (after tax) in GBP?	all categories	18,253
Geographic variables/ Country of Birth		all categories	**18,223**

References:

Gagné, Marylène, Jacques Forest, Maarten Vansteenkiste, Laurence Crevier-Braud, Anja van den Broeck, Ann Kristin Aspeli, Jenny Bellerose, et al. 2015. “The Multidimensional Work Motivation Scale: Validation Evidence in Seven Languages and Nine Countries.” *European Journal of Work and Organizational Psychology* 24 (2): 178–96. https://doi.org/10.1080/1359432X.2013.877892.