Significance of Task Significance in Online Marketplaces for Work

Yuqing Ren
University of Minnesota
chingren@umn.edu

Shawn Curley
University of Minnesota
curley@umn.edu

Abstract

Online marketplaces for work like Amazon Mechanical Turk facilitate the sourcing of low expertise tasks in a fast and cost effective way. In this study, we explore the impact of task significance on work quality by informing workers of the purpose of the task and who benefits from it. Results from a laboratory experiment and a field experiment showed that perceived task significance improved work quality, but only for participants who recalled the purpose statement. In contrast, increasing monetary payment by 50% had no impact on work quality. A majority of participants who received the purpose statement were not able to recall it. Further analysis showed worker attributes such as English ability and personality traits influenced the likelihood of recall whereas rich media format had no effects. Overall, our work highlights the promise of task significance as a way to motivate online workers and the challenge of promoting task significance online.

1. Introduction

Online marketplaces for work such as Amazon Mechanical Turk (MTurk) have emerged as a powerful new paradigm to accomplish low expertise work by crowdsourcing it to a large group of workers for small financial rewards [1]. Individuals and businesses use these platforms to accomplish a variety of small tasks such as image labeling, natural language processing, evaluation of search relevance, information verification, data cleaning, social science experiments and research data collection [2, 3, 4].

Despite the platforms' promise and utility, worker motivation and work quality remain as key challenges. There has been clear, consistent evidence of quality issues and gaming behaviors. In a survey of researchers who use MTurk to collect data, two thirds listed worker attentiveness or data quality as their greatest concern [5]. Several factors have been identified as causes to the quality issue such as low payment, lack of worker motivation and socialization, worker anonymity, various levels of work abilities [3, 6, 7].

Prior research has explored a variety of ways to reduce gaming behaviors and improve work quality by increasing payment, highlighting the meaning of the work, introducing qualification or screening tests, and providing feedback [3, 6, 7, 8]. Increasing monetary payment has increased work quantity but not work quality [7, 8]. Meaningfulness, operationalized as either nonprofit work or labeling tumor cells, has had mixed success in improving work quality. In one study [7], the nonprofit cover story increased the accuracy of counting malaria parasites in images. In another study [8], meaningfulness had no overall effects on workers across the globe and only marginal effects on American workers. The impact of meaningfulness also varies by worker location. In some studies, it reduced work output for workers from certain regions such as South Asia [7].

Prior studies have also explored ways to better screen and select workers to improve work quality. For instance, American workers have been shown to complete fewer tasks with higher accuracy than Indian workers [7]. Comparatively, older workers, female workers, professional and student workers are more likely to pass qualification tests than younger, male, and hourly workers [6].

In this paper, we aim to resolve the puzzle around the impact of meaningfulness on work quality in online labor markets. We explore two possible causes of the inconsistencies across studies: worker selection and the attention that workers pay to the instructions about meaningfulness. We conduct two experiments to examine perceived task significance as a prosocial motivation by informing workers of the purpose of the task and who would benefit from it. We also vary the levels of monetary payments partially to compare the effects of various motivations and partially to verify prior findings on the lack of effects of monetary payments. Participants performed proofreading tasks to fix errors in either Wikipedia articles or digitized books.

Experimental results showed that (1) task significance improved work quality, primarily when workers correctly recall and internalize the information about the purpose the task, (2) workers with greater English and cognitive abilities and certain personality traits such as agreeableness were better able to process the information and hence delivered higher quality work. Similar to prior literature, increasing monetary payment had no significant impact on work quality.

URI: https://hdl.handle.net/10125/79883
978-0-9981331-5-7

2. Theory and Hypotheses

2.1. Intrinsic and Extrinsic Motivations

Intrinsic motivation refers to the psychological state of wanting to do something "because it is inherently interesting or enjoyable" whereas extrinsic motivation refers to the psychological state of wanting to do something "because it leads to a separable outcome" [9, p. 55]. Decades of research in educational and other settings have shown that intrinsic motivation has robust, positive effects on task persistence and performance. While extrinsic motivation is also beneficial, it runs the risk of undermining intrinsic motivation [9].

Surveys and interviews of Amazon Mechanical Turk workers suggest that both intrinsic and extrinsic motivations can affect worker participation and their task choices [10]. Money has been shown as the most dominant motivating factor for MTurk workers [11, 12, 13], followed by enjoyment, free time, learning, and other factors. About 20-50% of workers report MTurk as their primary or secondary source of income**,** while 90% of tasks pay less than 10 cents [6, 12, 14], resulting in an hourly rate of $1 to $3 [4]**.** Research shows that it is often the perceived value of monetary payments, rather than the actual payments that matter [8, 15]. Higher payments increase the rate of signups and reduce dropout rates [7, 16], but the impact on work quality has been largely insignificant [8].

2.2. Prosocial Motivation

Prosocial motivation is the desire to expend effort to benefit other people [17]. Compared to intrinsic and extrinsic motivations, prosocial motivation has received less attention [18]. The concept is rooted in Hackman and Oldham's [19, 20] Job Characteristics Theory, which identifies five job characteristics that are likely to affect psychological states and work outcomes. One of the job characteristics is task significance, which is the degree to which the worker considers the job to be important or the job provides opportunities to positively impact the well-being of others [21].

Promoting task significance offers particular promise in the online labor marketplace where workers care about having an impact [24] but have limited awareness of the impact of their work. Yet, research studying the impact of meaningfulness in online marketplaces for work has led to mixed findings. In an image labeling task, MTurk workers who believed they were labeling tumor cells to assist medical researchers were more likely to watch a 3-minute video and label at least one image than workers who believed they were labeling objects of interest (80.6% vs. 76.5%) [25]. Nonetheless, the two groups did not differ on work quality. In another study, workers counted blood cells in an image, for either a nonprofit organization or a commercial firm. Workers who thought they were working for the nonprofit were consistently more accurate in their work [15].

Along the same line, meta-analyses of studies in traditional work settings show that task significance is strongly correlated with job satisfaction but only weakly correlated with job performance [22, 23]. Part of the reason may be that task significance usually is conceptualized as an objective attribute of the work itself. In contrast, studies that focus on workers' subjective perception of task significance [18, 21] have shown positive impact of task significance interventions on job performance. In addition, Grant's work [17, 21] reveals that several factors can moderate the effects of task significance, such as conscientiousness, prosocial values, and intrinsic motivation.

These results suggest two promising avenues for understanding meaningfulness in the online setting. First is the workers' perceptions of meaningfulness, especially their attention to meaningfulness as a task feature. Second is the role of worker attributes such as abilities and personalities that can enhance or suppress the effects of meaningfulness interventions.

We explore these possibilities in two experiments. Study 1 sets the stage by varying task significance and observing workers' attention to the message. We also manipulate monetary payments to benchmark and compare the effects of extrinsic and prosocial motivations. In Study 2, we further explore the attentional component of perceived task significance and ways to improve the chance for workers to process and internalize the task significance message.

3. Study 1 – Laboratory Experiment

Study 1 used a controlled setting for the initial investigation. Participants performed a proofreading task that mimicked Amazon Mechanical Turk. MTurk is a crowdsourcing platform launched in 2005. Requesters can post human intelligence tasks (HITs) and specify how many submissions they want and how much they will pay for each HIT. Workers search and choose to work on these HITs. The platform provides templates to post a broad range of HITs such as data collection (e.g., search for phone numbers of restaurants), data correction (e.g., check the spelling of search terms), image tagging, and research surveys. We chose proofreading as the task because: (1) it is representative of MTurk tasks in terms of nature and effort required, and (2) work quality can be assessed objectively by counting the number of errors fixed.

3.1. Methods

3.1.1. Participants. We recruited 173 participants through a behavioral research lab in a large Midwest university. The lab setting and college students as participants afforded us better control for the initial investigation. We posted the study as examining how people perform proofreading tasks in an online environment and included only participants who were proficient in English. A total of 166 participants provided valid demographic data. The majority were undergraduate (n=107) and graduate students (n=42) who attended the university. The mean age was 23.21 (SD = 7.33) and 39% of the participants were male.

3.1.2. Experimental task. Each HIT involved proofreading a 100-150 word paragraph, taken from featured articles at Wikipedia in the areas of business, sports, and music. We introduced three to seven random errors in each paragraph, such as switched letters, added or removed letters, and incorrect use of homophones. We shuffled articles in different areas so that each participant saw a mix of articles from all three areas. Here is an example paragraph with five errors highlighted in bold:

> In spite of popular belief, actuaries do not always **atempt** to predict aggregate future events. Often **there** work may relate to determining the cost of financial liabilities that have already **ocurred**, called retrospective reinsurance, or the development or re-pricing of new products. Actuaries also design and maintain products and systems. They are involved in **financal** reporting of companies' assets and liabilities. They must communicate complex concepts to clients who may not share their language or depth of knowledge. Actuaries work under a strict code of ethics that covers their communications and work products, but their clients may not adhere to those same standards when **enterpreting** the data or using it within different kinds of businesses.

Participants were instructed "to find all errors and typos and fix them so that the paragraph is accurate and free of error." They were informed that their payment depended on the number of paragraphs that they correctly fixed. We turned off the spell-checking function in the browser and asked participants not to use any external assistance while working on the task. All participants followed this instruction except one who was excluded from subsequent analysis.

3.1.3. Manipulations. We randomly assigned participants to one of four conditions within a 2 (purpose statement) x 2 (payment) between-subjects design. Participants in the task significance condition saw a paragraph titled "How the results of this HIT will be used and who will benefit," while participants in the control condition did not see the paragraph. Participants in the low payment condition were paid $0.20 for each paragraph versus $0.30 in the high payment condition. The task significance paragraph read as follows:

> These paragraphs come from articles on Wikipedia, a free online encyclopedia that anyone can edit. Wikipedia was launched in 2001 and is currently the largest and most popular general reference on the Internet. … By fixing typographical and spelling errors in these paragraphs, you will help improve the quality of these articles. Many Internet users who read and refer to these articles will benefit from your work.

3.1.4. Procedure. Each experimental session lasted one hour with up to six participants. After providing consent, participants sat at a computer, working individually. Conditions had been previously randomized and assigned to the computers. Participants completed a training task to learn all the steps needed to accept and complete a HIT. They then had 30 minutes to work on as many paragraphs as they wished.

Several studies have revealed that MTurk workers often multitask while working on the tasks [5]. We asked participants to perform a distraction task on paper to mimic the MTurk work setting for external validity. The distraction task is a cognitive ability test consisting of 30 questions (from http://www.wonderlic.com) that tested one's ability to understand instructions and solve problems. An example question is to identify the noun or verb in a sentence (e.g., Jill sets the plates on the table). Participants were told that it was up to them how to allocate their time between the two tasks.

After completing the tasks, participants completed a questionnaire about their experience of working on the task and their demographics. Participants were then debriefed and paid. The payment included $3 for showing up, $2 for completing the task on paper, and the amount earned in performing the spell-checking tasks, which ranged from $4 to $7.

3.1.5. Work quality measure. We measured work quality as the percentage of fixed errors. If a paragraph had five errors and a participant fixed three, work quality is 60%. A total of 2770 paragraphs were submitted. We wrote a Perl script to automatically calculate accuracy. We also had a research assistant manually code 1,055 paragraphs, and the manual coding and automatic coding were very consistent (r = 0.99). We analyzed the automatic coding.

3.2. Results

3.2.1. Manipulation checks. As a manipulation check, we asked about payment conditions and 95.8% of

participants correctly recalled their condition. We also asked participants to report their "best knowledge or best guess of how the fixed paragraphs will be used to benefit others." Surprisingly, the majority of the participants who had seen the purpose statement were not able to recall it. Only 28.4%, or 23 out of 81 participants, mentioned improving Wikipedia article quality. Most of the participants who did not see the task significance paragraph answered "don't know" or "have no idea." This supports the need for a distinction between task significance as an objective task feature and *perceived* task significance. We further pursued this idea in Section 3.2.3 and Study 2.

3.2.2. Effects of purpose statement on work quality. Table 1 shows the least square means across the four conditions. We conducted an ANOVA with task significance and payment as the independent variables and work quality as the dependent variable. The analysis showed no significant effects of seeing the purpose statement, $F(1, 169) = 2.3$, $p = 0.13$, increased payment, $F(1, 169) = 0.01$, $p = 0.92$, or their interaction, $F(1, 169) = 0.07$, $p = 0.8$.

Table 1. Effects of purpose statement and monetary payment on work quality

	Purpose Statement	
Payment	No Statement	Yes Statement
Low	0.646[a] (0.024)	0.615[a] (0.024)
High	0.654[a] (0.024)	0.611[a] (0.025)

Means with the same subscripts are not significantly different at $p < .05$ level.

3.2.3. Recall of purpose statement: Selection or motivation? As noted, most of the participants were not able to recall of the purpose statement. This suggests a disjoint between receiving the statement and their paying attention to the statement. To explore further, we split the participants in the purpose statement condition into two groups, those who correctly recalled the statement and those who did not. We ran an ANOVA with three levels: "no statement" meaning the participants were not exposed to the statement; "ignored" meaning the participants saw the statement but were not able to recall it, and "recalled" meaning participants received and recalled the statement. There is a significant positive effect of being able to recall the statement, $F(2, 167) = 7.33$, $p < 0.001$. As shown in Figure 1, the recalled group had the highest level of accuracy of 0.709, followed by no statement, 0.65, and the ignored group, 0.577. The differences between the ignored group and the other two groups were significant at the $p < 0.001$ level. The difference between the recalled and no statement groups was not significant ($p = 0.23$), possibly due to the small sample size of the recalled group ($N = 23$).

The next question is what might have led to the differences between the recalled and ignored groups. Was it a selection effect (e.g., filtering out participants with low cognitive ability), a motivation effect (e.g., motivating participants to exert effort), or a combination of the two? We conducted additional analyses to tease apart the two possible explanations. We measured cognitive ability by the percentage of questions that a participant had answered correctly in the distraction task. The score ranged between 23.3% and 100% with a mean of 75.9%. We analyzed cognitive ability as a mediator between recall of the purpose statement and work quality. If it were a pure selection effect we would expect the relationship between recall and work quality to be fully mediated by cognitive ability. If it were a combination of selection and motivation, we would expect a partial mediation, meaning the purpose statement provided additional motivation besides filtering out incapable workers.

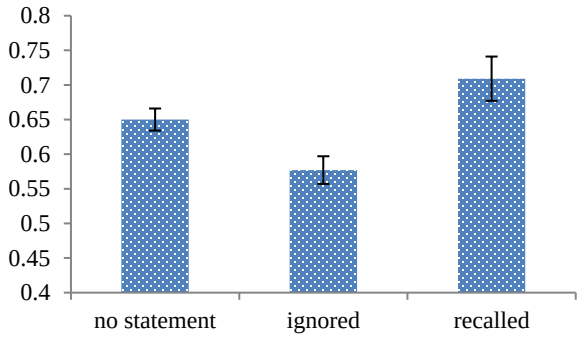

Figure 1: Effects of the recall of purpose statement on work quality

We tested the mediation effect, following the procedure described in Baron and Kenny [26]. First, we ran a logistic analysis to predict the likelihood of recall, and cognitive ability had a positive effect ($p = 0.001$). Participants in the recalled condition scored 0.842 in the cognitive ability test, and participants in the ignored condition scored 0.727. Second, we ran a regression analysis of work quality on statement recall and repeated the analysis with both recall and cognitive ability as the independent variables. Recall of purpose statement was significant in both analyses. Cognitive ability was also significant (0.376, $p < 0.01$). The inclusion of cognitive ability increased R-square from 0.08 to 0.15 and reduced the coefficient of recall from 0.132 ($p = 0.002$) to 0.09 ($p = 0.04$). It was a partial mediation, suggesting that the quality difference between the ignored and recalled groups was likely a combination of selection and motivation.

3.2.4. Interplay between task significance and intrinsic motivation. We measured intrinsic motivation with three items adapted from [17]: "I enjoyed performing the spell checking task," "I would return to MTURK to perform more of the spell checking task," and "I would perform the spell checking task even without the payment" (Cronbach's alpha = 0.6). We took their average to measure intrinsic motivation. Using the median value, we split the participants into two groups of low and high intrinsic motivation. An ANOVA showed a marginally significant, positive effect of statement recall, $F(1, 81) = 3.74$, $p = 0.06$, a positive effect of intrinsic motivation, $F(1, 81) = 5.4$, $p = 0.03$, and a marginally significant interaction between the two, $F(1, 81) = 5.6$, $p = 0.07$. The highest work quality was achieved with both the recall of purpose statement and high intrinsic motivation. This is consistent with findings from [17].

3.3. Study 1 Discussion

Similar to prior research, we found that monetary payment had no significant effects on work quality. Meanwhile, college students may not be representative of MTurk workers in their response to payments. The difference between 20 cents and 30 cents per HIT may mean more to MTurk workers than to college students.

The main finding of Study 1 is that the purpose statement to signal task significance had no impact on work quality, largely because most workers ignored or failed to register the information. Although the MTurk Best Practice Guide recommends that requesters inform workers of the purpose of the work, our results suggest that this practice needs to be implemented thoughtfully. Simply including a paragraph about how the work would benefit others is insufficient to motivate workers and improve work quality. Why did most participants ignore the purpose statement? What might have led to the differences in recalling the purpose statement?

We explored these questions in Study 2 by examining two factors that may affect the recall of the purpose statement: (1) the media format of the purpose statement and (2) worker attributes such as personality traits. Past research has shown that both the presentation of a message and the person receiving the message can influence the effectiveness of communicating the information [27].

In general, text is better remembered when it is supported with visual illustrations such as pictures [28, 29]. The facilitation of visual illustrations can be explained by dual coding theory [30, 31]. According to the theory, different types of information are processed in different cognitive subsystems, with words and sentences being processed in a verbal system and pictures being processed in an imagery system. Both forms of information can be kept simultaneously in working memory, and hence, they provide alternate paths to recall making it is easier to later retrieve the information [32]. Researchers have explored the use of multimedia such as visual or audio presentation to engage learners and reach a diverse audience [33]. We hence explore the use of video and audio formats, in addition to text, to communicate the purpose statement.

Research in text and multimedia comprehension has revealed individual differences in representational preferences and cognitive styles [32]. Several individual attributes may affect one's attention to and likelihood of, recalling a message. The first set of attributes are personality factors like extraversion, agreeableness, conscientiousness, neuroticism, and openness to experience [34]. Agreeableness, for instance, involves characteristics like altruism, nurturance, caring, and emotional support and often reflects one's prosocial motivation [17]. Individuals with high agreeableness tend to care more about others and are more likely to notice the purpose statement. Another factor is conscientiousness, which has been shown to be a reliable predictor of job and academic performance [35]. Prior knowledge may also matter because text comprehension requires adequate prior knowledge to be able to construct a mental model to comprehend the text [36]. Therefore, we investigate personality attributes and cognitive ability as potential individual differences that may affect the likelihood of recalling and internalizing the purpose statement.

4. Study 2 – Field Experiment

4.1. Methods

The experimental design and task are similar to Study 1. To increase perceived task significance, we changed the task from proofreading Wikipedia articles to proofreading e-books for underprivileged people. Workers were informed that the paragraphs are from out-of-print books that have been digitized and converted to electronic text using Optical Character Recognition (OCR) software. The software is not 100% accurate so errors can be made and need to be fixed before the books are released.

4.1.1. Participants. A total of 1,043 MTurk workers signed up to participate in the study and 1,005 completed at least one HIT. Among the 1,005 workers, 604 completed the survey, a response rate of 60%. According to the survey, 59.7% of the respondents were male, about a third were 18-24 years old and another third 25-34 years old, and 76% had a college or graduate degree. Workers were fairly evenly split between having full-time jobs, part-time jobs or being unemployed, and

78% of the respondents had an individual income of less than $20,000 per year. In general, workers in our study are representative of MTurk workers, with a slightly higher percentage of male workers and workers from non-US countries like India [4].

4.1.2. Manipulations. We manipulated three factors: task significance, media format, and monetary payment. There were three conditions of task significance: no statement, self-benefit statement, and purpose statement. Workers in the no statement condition saw only the instructions on how to perform the task. Workers in the self-benefit statement condition saw the instructions and a sentence about "how people who have worked on these tasks in the past have sometimes reported this as a good learning experience with interesting content." Workers in the purpose statement condition saw the instructions and a sentence saying, "By proofreading the text in this HIT, you will help preserve human knowledge and produce free e-books available to underprivileged people." We included the self-benefit condition to tease apart the effects of seeing any statement versus the purpose statement.

Within the purpose statement condition, we varied three media formats: plain text, a female-narrated video, and a male-narrated video. In the video conditions, the workers watched a captioned video with pictures of books and underprivileged people, voiced by a female or a male narrator. We varied two levels of payment: 15 cents versus 10 cents for each paragraph. Workers who completed the survey received an additional 50 cents.

Altogether, we had 5 (no statement, self-benefit, text, female video, male video) x 2 (10 vs. 15 cents) = 10 conditions manipulated using a between-subjects design. Similar to Study 1, work quality was measured as the average percentage of errors that were fixed.

4.1.3. Procedure. We posted the HIT on MTurk. After workers accepted the HIT, a built-in link automatically re-directed them to a university-hosted server. This enabled us to randomly assign the workers to experimental conditions and to avoid having one worker in multiple conditions.

The paragraphs were taken from Project Gutenberg (http://www.gutenberg.org), a site that organizes volunteers to collectively edit digitized versions of old books that are no longer bound by copyright in order to create free e-books. We gathered pages from five books on topics related to public speaking, ornithology, and food preparation. The paragraphs were taken from scans of old books, many of which are blurry and hard to read. Figure 2 shows our instructions to workers with the scanned page on the left and the converted text on the right. Workers were told the errors occurred because we used OCR software to convert the scanned image to text. Their goal was to find and fix all the errors so that the text matched the original text in the scanned image.

Compared to the task in Study 1, this modified task is less dependent upon English proficiency. Workers can simply compare the text in the text box (on the Right) to the original image (on the Left) to find and fix the errors. After eight paragraphs, a link began to appear under the task, informing the workers that they had the

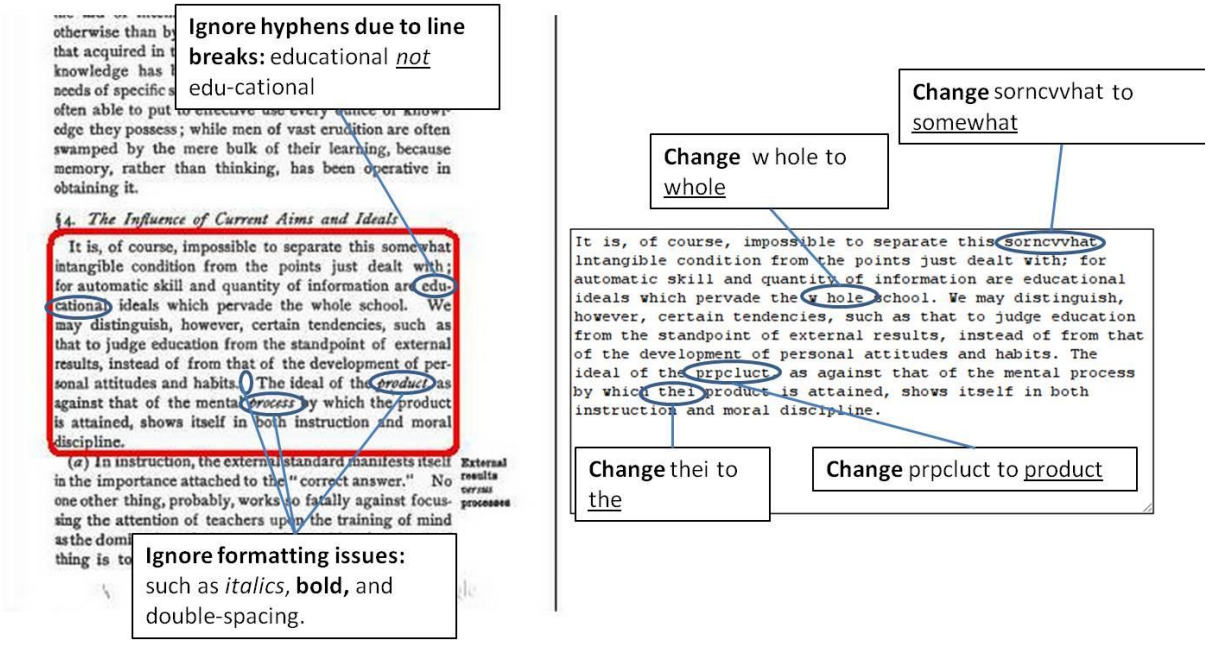

Figure 2: Instructions on how to proofread a paragraph

option to stop the proofreading task anytime and take a short survey for additional payment, or they could wait until all HITs were completed. The survey included questions about recall of the purpose statement, perceived task significance, worker demographics and the Big-Five Factor personality scale [37].

4.2. Results

4.2.1. Effects of purpose statement on work quality. Table 2 shows the ANOVA results. There is a positive effect of purpose statement, $F(2, 995) = 5.76$, $p = 0.003$, no significant effect of payment, $F(1, 995) = 0.15$, $p = 0.7$, and no interaction between the two, $F(2, 995) = 0.5$, $p = 0.61$. Pairwise comparisons show that workers in the purpose statement and self-benefit statement conditions outperformed workers in the no statement condition (89.6% and 90.4% vs. 86.9%).

Table 2. Effects of purpose statement and monetary payment on work quality

	Purpose Statement		
Payment	No Statement	Self-Benefit Statement	Purpose Statement
Low	0.877^a (0.011)	0.902^b (0.01)	0.896^b (0.006)
High	0.862^a (0.012)	0.906^b (0.013)	0.897^b (0.006)

Means with the same subscripts are not significantly different at $p < .05$ level.

4.2.2. Perceived task significance. Similar to Study 1, we asked workers "Who do you think will benefit from your work on proofreading the paragraphs?" and "How will they benefit from your work on proofreading the paragraphs?" We use the recall data as a behavioral indicator of perceived task significance. Two research assistants coded the responses separately and reached agreement on their coding. Of the 309 survey respondents who received the purpose statement, 116 or about 37.5% were able to recall it, slightly higher than the 28.4% in Study 1. We labeled these workers as the "recalled" group and the others as the "ignored" group.

ANOVA showed significant differences in work quality across the four groups of no statement, self-benefit statement, ignored, and recalled, $F(3, 997) = 7.86$, $p < 0.001$. As shown in Figure 3, workers who correctly recalled the purpose statement performed at the highest level compared to those in the other conditions (92.5% vs. 86.9% no statement, 90.4% self-benefit, 88.8% ignored).

Being able to recall the purpose statement does not necessarily mean that workers have internalized the message or changed perceptions of task significance. As a more direct measure of perceived meaningfulness, we adapted a scale from [21] to measure perceived task significance as: (1) the task provides opportunities to improve the welfare of others, (2) some people will be positively affected by how well I perform the task, (3) the task provides opportunities to have positive impact on others, and (4) the task seems trivial and does not have positive impact on others (reversed). We conducted an exploratory factor analysis, and all but the last item loaded on a single factor. Thus, we dropped the last item and averaged the first three items.

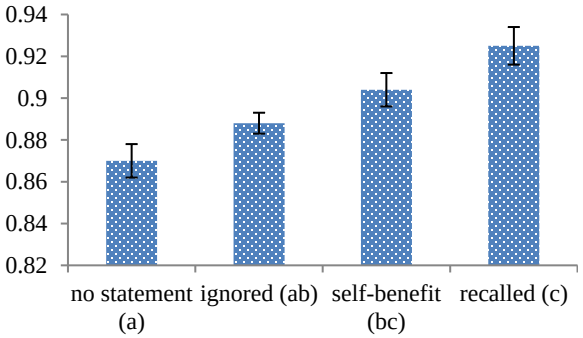

Figure 3: Effects of recall of purpose statement on work quality

ANOVA showed significant differences in perceived task significance across conditions, $F (3, 600) = 14.44$, $p < 0.001$. As shown in Figure 4, workers in the recalled condition reported the highest level of perceived task significance compared to workers in the other conditions (4.298 vs. 3.959 for ignored, 3.835 for self-benefit statement, and 3.79 for no statement).

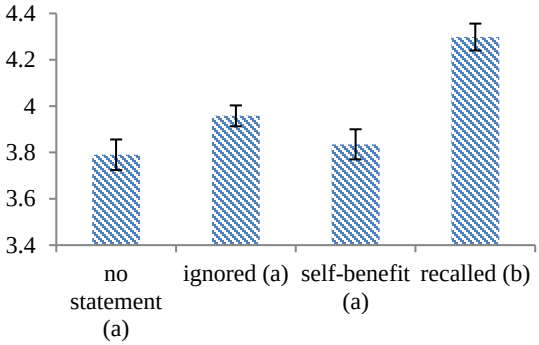

Figure 4: Effects of recall of purpose statement on perceived task significance

We then tested perceived task significance as a mediator between the purpose statement manipulation and work quality, following the procedure described in [25]. We first converted the four conditions to three indicator variables – self-benefit, ignored, and recalled – using no statement as the base. (1) We regressed work

quality on the three indicator variables and found a significant positive effect of all three (0.042, p = .002 for self-benefit, 0.025, p = .03 for ignored, and 0.046, p < .001for recalled). (2) We regressed perceived task significance on the three indicator variables and found a significant positive effect on only recalled (0.357, p < .001), meaning only workers in the recalled condition reported a higher level of perceived task significance. (3) We regressed work quality on the three indicator variables and perceived task significance. Perceived task significance was significant and positive (0.016, p = .01) and the effect of recalled dropped from 0.046 (p < .001) to 0.036 (p = .008). Hence, perceived task significance partially mediates the effects of the purpose statement on work quality.

4.2.3. Recall of the purpose statement: Individual attributes or media format? Workers self-reported their age, gender, education, English ability, individual income, and experience with MTurk. We used the Big Five Personality dimensions [37]. We asked participants to indicate the extent to which a list of attributes could be used to describe them, with six items for each of the five factors: conscientiousness, agreeableness, neuroticism, openness, and extraversion. Confirmatory factor analysis show most items loaded as expected.

We ran logistic regressions to predict whether a worker was able to recall the purpose statement. Table 3 shows the main results. Several individual attributes emerged as significant. On average, workers were more likely to correctly recall the message if they rated their English ability as "very well" or "extremely well," had heard of e-books, or had an individual income of more than $20,000. In addition, workers with high agreeableness were more likely to recall. Worker age, gender, and tenure with MTurk had no significant impact on the likelihood of recall.

Contrary to our expectations, media format had no significant impact on the likelihood of recall. Neither video condition increased the likelihood of recall than the text format. 46 or 38.66% of participants in the text condition recalled the statement, 38 or 37.25% of participants in the female-narrated video condition recalled it, and 32 or 36.36% of participants in the male-narrated video condition recalled it.

4.2.4. Interactions among different types of motivations. We included an open-ended question in the survey asking workers to provide additional comments about their experiences. Surprisingly, a good number of workers wrote about how much they enjoyed performing the HIT, so we manually coded whether a worker had mentioned fun or enjoyment in the open-ended response as a proxy for intrinsic motivation. Among the 603 survey respondents, 67 or 11.1% mentioned enjoyment. ANOVA results suggest a marginally significant effect of enjoyment on work quality, F(1, 601) = 3.42, p = 0.06. Workers who mentioned enjoyment performed with higher accuracy than those who did not report any enjoyment (92.5% versus 89.9%). We also found an interesting interaction between enjoyment and purpose statement, F(1, 484) = 2.66, p = 0.1). Workers in the purpose statement condition who also mentioned enjoyment performed at the highest level, higher than all the other conditions.

Table 3. Predicting the likelihood of purpose statement recall

Variables	Model 1	Model 2	Model 3
Intercept	-1.574** (0.569)	-3.632** (1.129)	-3.651** (1.134)
Newcomer (< 6 months)	0.142 (0.153)	0.127 (0.154)	0.129 (0.155)
Young Age (<24 years old)	-0.008 (0.145)	-0.011 (0.155)	-0.015 (0.155)
Female	0.118 (0.133)	0.138 (0.135)	0.136 (0.136)
High English Ability	**0.51**** (0.186)	**0.47*** (0.19)	**0.465*** (0.19)
Heard Of Ebooks	**1.251*** (0.523)	**1.256*** (0.523)	**1.272*** (0.524)
Income (< 20K)	**-0.657**** (0.22)	**-0.689**** (0.225)	**-0.698**** (0.227)
Income (20-40K)	-0.046 (0.29)	-0.035 (0.294)	-0.03 (0.296)
Agreeableness		**0.507*** (0.25)	**0.513*** (0.251)
Conscientiousness		0.04 (0.182)	0.033 (0.183)
Female Video			0.072 (0.19)
Male Video			-0.111 (0.193)
N	289	288	288
-2 Log Likelihood	347.1	341.1	340.8

Note: The base for Income is > $40K. The base for Condition is Text. $^{**}p < 0.01$, $^{*}p < 0.05$, $^{+}p < 0.1$.

4.3. Study 2 Discussion

Study 2 provides further evidence for the positive effects of a purpose statement and perceived task significance on work quality. Compared to college students in Study 1, a higher proportion of MTurk workers correctly recalled the purpose statement (37.5% vs. 28.4%). Similar to Study 1, we did not find any significant effects of monetary payments on work quality. We further found that individual attributes, not media format, predicted workers' likelihood of recalling

the purpose statement. On average, workers with high English ability, who had heard of e-books, with more than $20K annual income, and with a higher agreeableness personality were more likely to recall the statement. Delivering the statement in video format with either female or male narrated voice had no significant impact on the likelihood of recall.

Several things are worth noting. The first is the lack of any effects of media format. According to dual coding theory, better retention and recall occur when different types of information enter into one's memory through different sensory systems, e.g., reading about a geometry theorem and seeing a visual illustration of it [32]. In our experiment, the same information appeared in both text and video format. As a result, the use of video might have contributed to an overload of the mental systems, instead of creating a synergistic integration between the two media formats.

The second finding worth discussing is the positive impact of the self-benefit statement. Workers in the self-benefit condition performed almost as well as workers who correctly recalled the purpose statement. The result suggests an alternative promising direction to improve work quality by highlighting self-benefit.

5. General Discussion

We explore the concept of perceived task significance and its impact in online marketplaces for work. Results from two experiments demonstrated the promise of task significance in motivating online workers, and also revealed unexpected challenges in promoting task significance in online environments. About two-thirds of participants exposed to the purpose statement failed to recall or internalize the message. Those who were able to recall the statement performed at a level of accuracy that is about 6% higher than those who did not receive the statement. We believe these findings are theoretically and practically significant. For a project of 10K image tagging, it would mean about 600 more images being properly labeled.

In our efforts to explore factors that predict purpose statement recall, we found that a rich media format had no significant impact on workers' recall. The finding is counterintuitive and has important implications for the design of instructions in online marketplaces for work. Simply converting textual information to video displays is insufficient to improve workers' attentiveness. Future research should explore other ways to communicate the meaning of work in online labor markets. Instead, we found that worker attributes affected recall likelihood, which shows the promise of using individual ability and personality traits to screen and recruit workers.

In terms of motivating workers, our study combined with prior research suggests that monetary payment had limited power in improving work quality. Instead, there was a synergy between task significance and enjoyment as intrinsic motivation. Workers performed at the highest level of quality when both intrinsic and prosocial motivations were high.

Our study also provided methodological insights on the comparison of college students and MTurk workers as research participants. Previous research has provided mixed insights. Some studies found that MTurk workers paid less attention to instructions and were less likely to answer questions correctly than college students [38] whereas other studies found that MTurk workers paid greater attention to instructions. We found that MTurk workers outperformed college students in both their attentiveness and work quality (37.5% vs. 28.4% in recall and 92.5% vs. 72% in work quality) providing support for the use of MTurk for research purposes.

Despite our best efforts, there are limitations. We only experimented with one type of task – proofreading paragraphs – which requires English literacy and cognitive ability. We recruited MTurk workers from multiple countries, primarily the US and India, which makes it difficult to separate the effects of national cultures and personal attributes. In terms of monetary payments, we experimented with a small range (20 vs. 30 cents in Study 1 and 10 vs. 15 cents in Study 2). It is possible that the impact of payment is non-linear and that, given our payments are higher than the average MTurk payment, our ranges fell in the leveling-off region where the effects are asymptotic. As online labor markets continue to grow to be an indispensable part of our global economy, more research is needed to understand and inform the practice of how to effectively recruit and motivate online workers.

7. References

[1] J. Howe, Crowdsourcing: Why the Power of the Crowd Is Driving the Future of Business. Crown Business: New York, 2008.

[2] D. Feng, S. Besana, and R. Zajac, "Acquiring High Quality Non-Expert Knowledge from On-demand Workforce", In Proceedings of the 2009 Workshop on The People's Web Meets NLP, Suntec, Singapore, pp. 51-56, 2009.

[3] A. Kittur, E.H. Chi, and B. Suh, "Crowdsourcing User Studies with Mechanical Turk", In Proceedings of the Conference on Human Factors in Computing Systems, Florence, Italy, 2008.

[4] J. Ross, L. Irani, M.S. Silberman, A. Zaldivar, and B. Tomlinson, "Who are the Crowdworkers? Shifting Demographics in Amazon Mechanical Turk", In Proceedings of Conference on Human Factors in Computing Systems, Atlanta, GA, 2010.

[5] J. Chandler, P. Mueller, and G. Paolacci, "Nonnaivete among Amazon Mechanical Turk workers: Consequences and solutions for behavioral researchers", Behavior Research Methods, 46(1), 112-130, 2014.

[6] J.S. Downs, M. B. Holbrook, S. Sheng, and L.F. Cranor, "Are Your Participants Gaming the System? Screening Mechanical Turk Workers", In Proceedings of the Conference on Human factors in Computing Systems, Atlanta, GA, 2010.
[7] J. Rogstadius, V. Kostakos, A. Kittur, B. Smus, J. Laredo, and M. Vukiovic, "An Assessment of Intrinsic and Extrinsic Motivation on Task Performance in Crowdsourcing Markets," In Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media, pp. 321-328, 2011.
[8] W.A. Mason, and D.J. Watts, "Financial Incentives and the Performance of Crowds", In Proceedings of ACM Workshop on Human Computation, pp. 77-85, 2009.
[9] R.M. Ryan and E.L. Deci, "Intrinsic and Extrinsic Motivations: Classic Definitions and New Directions", Contemporary Educational Psychology, 25, pp. 54-67, 2000.
[10] S. Moussawi and M. Koufaris, "Working on Low-Paid Micro-Task Crowdsourcing Platforms: An Existence, Relatedness and Growth View," In Proceedings of the 36th International Conference on Information Systems, Fort Worth, TX, 2015.
[11] J. Antin, and A. Shaw, "Social Desirability Bias and Self-Reports of Motivation: A Study of Amazon Mechanical Turk in the US and India," In Proceedings of the ACM 2011 Conference on Computer Supported Cooperative Work, Hangzhou, China, 2011.
[12] P. Ipeirotis, "Demographics of Mechanical Turk", Working Paper, New York University, 2010.
[13] N. Kaufmann, and T. Schulze, "More than Fun and Money. Worker Motivation in Crowdsourcing – A study on Mechanical Turk", In Proceedings of the 17th Americas Conference on Information Systems, Detroit, MI, 2011.
[14] P. Ipeirotis, "Analyzing the Amazon Mechanical Turk Marketplace", Working Paper. New York University, 2010.
[15] S. Moussawi and M. Koufaris, "The Crowd on the Assembly Line: Designing Tasks for a Better Crowdsourcing Experience," In Proceedings of the 34th International Conference on Information Systems, Milan, Italy, 2013.
[16] M.J.C. Crump, J.V. McDonnell, and T.M. Gureckis, "Evaluating Amazon's Mechanical Turk as a Tool for Experimental Behavioral Research", PLOS ONE, 8(3), pp. 1-18, 2013.
[17] A.M. Grant, "Does Intrinsic Motivation Fuel the Prosocial Fire? Motivational Synergy in Predicting Persistence, Performance, and Productivity", Journal of Applied Psychology, 93(1), pp. 48-58, 2008.
[18] A.M. Grant "Relational Job Design and the Motivation to Make a Prosocial Difference", Academy of Management Review, 32(2), pp. 393-417, 2007.
[19] J.R. Hackman, and G.R. Oldham, "Motivation through the Design of Work: Test of a Theory", Organizational Behavior and Human Performance, 16(2), pp. 250-279, 1976.
[20] J.R. Hackman, and G.R. Oldham, Work Redesign. Addison-Wesley Publishing, 1980.
[21] A.M. Grant, "The Significance of Task Significance: Job Performance Effects, Relational Mechanisms, and Boundary Conditions", Journal of Applied Psychology, 93(1), pp. 108-124, 2008.
[22] Y. Fried, and G.R, Ferris, "The Validity of the Job Characteristics Model: A Review and Meta-analysis", Personnel Psychology (40), pp. 287-322, 1987.
[23] S. E. Humphrey, J. D. Nahrgang, and F.P. Morgeson, "Integrating Motivational, Social, and Contextual Work Design Features: A Meta-Analytic Summary and Theoretical Extension of the Work Design Literature", Journal of Applied Psychology, 92(5), pp. 1332-1356, 2007.
[24] X.N. Deng, K.D. Joshi, and R.D. Gallier, "The Duality of Empowerment and Marginalization in Microtask Crowdsourcing: Giving Voice to the Less Powerful through Value Sensitive Design", MIS Quarterly 40(2), 279-302, 2016.
[25] D. Chandler and A. Kapelner, "Breaking Monotony with Meaning: Motivation in Crowdsourcing Markets", Working Paper. University of Chicago, 2012.
[26] R.M. Baron, and D.A. Kenny, "The Moderator-Mediator Variable Distinction in Social Psychological Research: Conceptual, Strategic, and Statistical Considerations", Journal of Personality and Social Psychology, 51(6), pp. 1173–1182, 1986.
[27] S. Kalyuga, P. Chandler, and J. Sweller, "Incorporating Learner Experience into the Design of Multimedia Instruction", Journal of Educational Psychology, 92(1), pp. 126-136, 2000.
[28] H. W. Levie and R. Lentz, "Effects of Text Illustrations: A Review of Research", Educational Communication and Technology, 30: pp. 195–232, 1982.
[29] J.R. Levin, G.J. Anglin, and R.N. Carney, "On Empirically Validating Functions of Pictures in Prose", In Willows, D. M., and Houghton, H. A. (eds.), The Psychology of Illustration, Vol. 1, Springer, New York, pp. 51–86, 1987.
[30] J.M. Clark and A. Paivio, "Dual Coding Theory and Education", Educational Psychology Review, 3: pp. 149–210, 1991.
[31] A. Paivio, Mental Representations: A Dual Coding Approach, Oxford University Press, Oxford, 1986.
[32] W. Schnotz, "Towards an Integrated View of Learning from Text and Visual Displays", Educational Psychology Review, 14(1), pp. 101-120, 2002.
[33] T.F. Hawk and A.J. Shah, "Using Learning Style Instruments to Enhance Student Learning. Decision Sciences", Journal of Innovative Education, 5(1), 1-19, 2007.
[34] R.R. McCrae and O.P. John, "An Introduction to the Five-Factor Model and Its Applications", Journal of Personality, 60(2), pp. 175-215, 1992.
[35] O.P. John and S. Srivastava, "The Big-Five Trait Taxonomy: History, Measurements, and Theoretical Perspectives", In L. Pervin and O.P. John (Eds), Handbook of Personality: Theory and Research (2nd ed.). New York: Guilford, 1999.
[36] D.S. McNamara, E. Kintsch, N.B. Songer, and W. Kintsch, "Are Good Texts Always Better? Interactions of Text Coherence, Background Knowledge, and Levels of Understanding in Learning from Text", Cognition and Instructions, 14(1), pp. 1-43, 1996.
[37] M.R. Barrick and M.K. Mount, "The Big Five Personality Dimensions and Job Performance: A Meta-Analysis", Personnel Psychology, 44, pp. 1-26, 1991.
[38] J.K. Goodman, C. E. Cryder, and A. Cheema, "Data Collection in a Flat World: The Strengths and Weaknesses of Mechanical Turk Samples. Journal of Behavioral Decision Making, 26, pp. 213-224, 2013.