Proceedings of the Fifth Conference on
Human Computation and Crowdsourcing
(HCOMP 2017)

"But You Promised": Methods to Improve Crowd Engagement In Non-Ground Truth Tasks

Avshalom Elmalech
Harvard University, USA
elmalech@gmail.com

Barbara J. Grosz
Harvard University, USA
grosz@eecs.harvard.edu

Abstract

Crowdsourcing platforms were initially designed to recruit people to perform tasks that were simple cognitively but difficult for computers. One challenge in these settings is to identify an incentive mechanism for motivating workers to complete tasks and do high-quality work. Previous research has studied the use of financial incentive mechanisms and social comparison as motivators. These mechanisms can only be applied to ground truth tasks, tasks for which there is an objective performance scale. In this paper, we define and compare three innovative methods for improving worker engagement on non-ground truth tasks drawing on a psychological theory of commitment. The three methods are similar in asking participants to promise they will complete a task, but they differ in terms of how the commitment is made. In the first method, participants commit by signing a contract; in the second, by listening to a recording; in the third, by recording a personal commitment. The last two methods significantly improved the task completion rate when compared to two baseline conditions. The methods we propose can be implemented simply, can be used for any task, and do not affect participants' behavior other than by improving their engagement.

Introduction

Over the past decade there has been constant growth in participation in online crowdsourcing platforms. One main challenge all crowdsourcing platforms share is that of motivating participants. Some platforms (e.g., Amazon Mechanical Turk (http://www.mturk.com/)) use financial compensation mechanisms for recruiting people. Other platforms (e.g., Zooniverse (Borne and Team 2011)) use people's desire to contribute to science as a means of motivating them to carry out tasks that will further scientific research (without offering them monetary compensation). The lab in the wild platform (Reinecke and Gajos 2015) relies on people's curiosity to motivate them to participate in the tasks. In this work, we introduce methods to improve crowd engagement in tasks on financial-based platforms. In these platforms, the use of "simple" extrinsic incentives to improve work quality (e.g., higher payment for completing a task) is not sufficient (Yin and Chen 2015; Radanovic and Faltings 2016; Singla and Krause 2013).

The lack of commitment to doing high-quality work on those platforms may be attributed to the absence of a long-term relationship between workers and employers. Financial-based crowdsourcing platforms do not encourage or facilitate such long term relationships. In traditional labor markets, where there is a long-term employee-employer relationship, a "psychological contract" forms between the two parties (Robinson 1996). This psychological contract incentivizes commitment and a willingness to engage in organizational citizenship behavior (Coyle-Shapiro and Kessler 2000). In some crowdsourcing platforms, the lack of relationship between the employees and the employers puts the employees in a state of constant tension. On one hand, the employees know that their employer expects them to be fully engaged in the task at hand - that is what they are being paid for. On the other hand, employees attempt to maximize their revenue which makes the dominant strategy to complete a task as quickly as possible, resulting in low-quality results). The main benefit for employers in using these platforms is the large, diverse and stable pool of workers they offer. In addition, the compensation for completing a task is relatively low compared to traditional labor markets.

Several approaches for enhancing worker engagement in ground truth (GT)-based tasks (there is an objective performance scale) have been studied. Some research examines the use of financial compensation mechanisms which are based on the workers' performance (Yin and Chen 2015; Mason and Watts 2010; Gao *et al.* 2012; Feng *et al.* 2014). Another line of research suggests the use of intrinsic behavioral factors to motivate workers (Law *et al.* 2016; Kaufmann *et al.* 2011).

Keeping workers motivated on tasks for which there is no ground truth (no right or wrong answer) is even more difficult. For non-ground truth (NGT) tasks, the traditional "carrot and stick" approach cannot be applied because in situations in which there is no benchmark for work quality, it is very difficult (and sometime impossible) to assess work quality. Workers participating in this type of task are sometimes aware that there is no quality metric, which can affect the way they carry it out. To illustrate this problem, imagine a scenario where a company is interested in launching a new website. To get feedback about the design and the quality of interaction with the new website, the company decides to use workers from an online labor market. The workers are asked to share their personal experience, so they know that the quality of their work is not measurable. They are faced with the following dilemma: should they do the right thing and invest the time to truly interact with the webpage before supplying their answers, or should they simply supply random answers and complete the task in less time. Imposing compensation mechanisms based on the similarity of one worker's answer to that of another is dangerous because workers may not indicate their true preferences, and instead attempt to list preferences they believe are common to most of the workers. The use of crowdworkers for NGT tasks is very common (Hossfeld *et al.* 2014; Loni *et al.* 2014; Larson *et al.* 2014; Soleymani *et al.* 2012), it is therefore important to address the crowd engagement problem for NGT tasks.

In this paper, we introduce three methods for improving worker engagement in NGT tasks. The inspiration for our methods came from the "commitment and consistency" theory of Cialdini and Garde (Cialdini and Garde 1987). According to this theory, the desire for consistency is embedded in human nature. Inconsistency is commonly thought to be an undesirable personality trait. The desire for consistency is external and internal: we desire to be perceived as consistent human beings by society and by *ourselves*. According to this theory, the desire to be and to look consistent is triggered when we commit to a certain action. When a commitment is made, it is likely to be met even in situations where it does not align with one's best interest. The "commitment and consistency" theory introduces the effect of commitment on people's actions, but it does not suggest what form of commitment is most effective.

In this paper, we compare the efficiency of three ways crowdworkers may commit to doing their best in completing a task:

- "Commitment by signature": The workers are required to confirm they will complete the task exactly the way it is stated in the task description by checking a box.
- "Commitment by listening": The workers are required to confirm they will complete the task exactly the way it is stated in the task description by listening to a recording stating that they will.
- "Commitment by speech": The workers are required to confirm they will complete the task exactly the way it is stated in the task description by recording themselves stating they will.

The results reported in this paper indicate that the most effective forms of commitment are the commitment by speech and the commitment by listening. As for commitment by signature, the improvement was not statistically significant when compared to two baseline conditions.

The main contribution of this work is two-fold:

- Proposing domain independent mechanisms (i.e., the mechanisms do not involve any modification to the domain in which they apply) for improving crowd engagement in NGT tasks and demonstrating the efficiency of these mechanisms.
- Validating the non-desirable behavior of crowdworkers in NGT tasks and demonstrating that this behavior is not due to poor attention span.

Related Work

Crowdsourcing and Motivation

The past decade has seen significant growth in crowdsourcing applications. There are two types of tasks in which crowdsourcing is used. One type is the simple or independent task. For example, labeling data for machine vision (Patterson and Hays 2012; Patterson *et al.* 2014; Raykar *et al.* 2010) and natural language processing (Snow *et al.* 2008). The second type is the complex and higher-order cognition task (Kittur *et al.* 2013). Much of the research dealing with the second type of task has dealt with the question of how to use decomposition techniques to break down a complex task into independent and simple micro tasks (Bernstein *et al.* 2015; Bigham *et al.* 2010; Kim *et al.* 2014; Kittur *et al.* 2011; 2012; Kulkarni *et al.* 2011; Lasecki *et al.* 2013; Little *et al.* 2010).

The crowdsourcing model, although novel in some important respects, retains a chalenge for labor marker researchers key question that has captured the attention of researchers: *How to motivate workers and improve worker engagement so they perform higher quality work*. This challenge is of great importance, especially given the evidence that some workers are primarily interested in producing quick rather than quality labor in order to optimize their efficiency and, in turn, earn more money (Laws *et al.* 2011; Akkaya *et al.* 2010).

Approaches to Motivating Crowdworkers

As the user motivation problem is very important for the future of financial-based crowdsourcing platforms, various methods have been proposed to address this problem. We review three main approaches and explain the reasons they are not applicable to NGT tasks. The first line of work suggests using financial incentives to motivate workers (Mason and Watts 2010; Kittur *et al.* 2008; Yin *et al.* 2014; Liu and Chen 2016). Generously rewarding workers upon completing a task was found inadequate in ensuring better work quality. In fact, it has been shown that even at low compensation rates, payment levels do not appear to affect work quality but only the number of tasks workers complete (Mason and Watts 2010). Moreover, it was shown that financial incentives can be used to trade quality for speed (Mao *et al.* 2013). The alternative to fixed payment upon completion of a task is a reward based on performance. Different variants of performance-based incentives have been studied. In one variant, workers are rewarded based on "relative performance" (also known as "crowd-sourcing contests" (Gao *et al.* 2012; Feng *et al.* 2014)). Namely, only the worker who performed best receives a reward. In another variant, workers are rewarded based on "absolute satisfying performance", i.e., based on pre-specified criteria (Ho *et al.* 2015). According to this method, each worker who meets the pre-specified criteria receives the reward regardless of the performance of the other workers. These methods do not apply to NGT tasks since performance cannot be measured.

The second line of work suggests the use of gold-standard data to monitor work quality (Biewald 2011; Drapeau *et al.* 2016). This approach uses fear of rejection to motivate workers. The workers are told that the employer has the correct answers to some of the questions that appear in the task, but the workers do not know which questions these are. In addition, the workers are told that if they don't answer these particular questions correctly their entire work will be rejected and they will not get paid. This approach, although promising, cannot be used in non-ground truth tasks since there are no right or wrong answers. Therefore gold-standard data cannot be embedded in the task.

The third line of work suggests that intrinsic motivations such as enjoyment (Kaufmann *et al.* 2011), curiosity for new knowledge (Law *et al.* 2016) or interest in contributing to a certain scientific discipline (Mao *et al.* 2013) can incentivize crowd-sourced workers. Although promising, for some tasks such intrinsic motivational drivers are not feasible.

Motivating Crowd-Sourced Workers using Their Peers

The research question this paper addresses —motivating crowd-sourced workers in *non-ground truth* tasks, has been investigated in the past. A peer prediction mechanism (or a variant of this mechanism) has been suggested as a means to address the worker motivation problem (Kamar and Horvitz 2012; Witkowski and Parkes 2012; Dasgupta and Ghosh 2013; Huang and Fu 2013b; 2012). The general idea behind this mechanism is that each task/question is given to several workers, and the compensation for answering the question is based on the number of workers who gave the same answer. One problem with this mechanism is that it may lead to consensus bias. The mechanism may encourage participants who know their opinion is of a minority group to change their answer to one they believe is shared by the majority group in order to maximize their compensation. A Bayesian truth mechanism (Prelec 2004) was suggested to address the consensus bias issue. The mechanism requires workers to supply, along with their answer, a prediction of the entire populations answers. There are several limitations to deploying this method in crowdsourcing labor markets. First, it requires workers to predict their peers answers. In many tasks workers do not have this information. Second, the underlying assumption of the mechanism is that the players are utility maximizers. This assumption is problematic, since it has been shown in many domains that people are rationally bounded (Simon 1982; Kahneman 2000; Shin and Ariely 2004).

Using any form of monetary incentive for eliciting a preference may result in participation bias. It has been shown (Hsieh and Kocielnik 2016) that in online labor markets, monetary incentives affect the kind of workers who are willing to take a job. In NGT tasks where it is important to capture the preferences of the entire population, this bias is likely to be unacceptable to employers.

An interesting line of research suggests the use of social facilitation as a means of motivating crowd-sourced workers (Huang and Fu 2013a). In this line of work, the labor performed by workers is compared to the labor performed by their peers; however the compensation for a certain worker does not depend on the work performed by other workers. The incentives in this case are social, therefore this method is less likely to suffer from participation bias. Although promising, this mechanism requires the overhead of pairing together workers and asking them to share private information about themselves.

Commitment and Consistency

The methods reported in this paper were inspired by commitment and consistency theory (Cialdini and Garde 1987). This theory has influenced disciplines besides marketing. In one psychological experiment, (Schwartz *et al.* 2014) it was shown that when the theory is applied properly it can promote positive health-related choices. In that study, participants' pre-commitment to a healthier life style (i.e., purchasing healthier food) improved their choices. Another example of the use of this theory is the "stickk" goal setting application (http://www.stickk.com). The application assists people in achieving their goals by asking them to set goals and put some money aside, the money will be returned to them only if the goals are achieved. People sign a contract obligating them to achieve their personal goals. The creators of this application stress that the commitment contract this application offers is the key to its success.

Another relevant piece of psychological research deals with "self-prophecy effect" (Greenwald *et al.* 1987; Sherman 1980). When people are asked to predict the probability that they will take a certain action, their behavior is more likely to confirm that prediction. In one study, (Sherman 1980) participants who were asked to predict the likelihood that they would contribute to a charity were more likely to do so when given the opportunity than participants who were not asked to predict their behavior. In another study, (Greenwald *et al.* 1987) the proportion of participants who answered "yes" when asked if they would vote in an upcoming election, and went on to do so, exceeded the proportion of participants who voted without being asked to predict their behavior.

Experimental Infrastructure

Figure 1 illustrates the experimental work-flow. Participants first got instructions, namely the task description which was to check 20 images of artificial grass and indicate which one looked most natural. Participants were instructed to check all 20 images before deciding; this specific requirement was highlighted. On the second, third and fourth steps of the experiment, participants were required to record themselves, listen to a recording and read one line of text. The content of what they were required to record, listen to, and read varied based on the experimental conditions. In the next section we will describe the exact content in more detail. The fifth step was the actual task, the artificial grass survey. On the final step of the experiment, participants were required to supply demographic information (gender, age and location).

Figure 2 shows an example of the actual grass survey task. There were 20 links to web pages containing the artificial grass images. Participants were required to select one answer to complete the survey. The 20 links were in a text format and not hyperlinks (e.g, clicking on the links would not direct the worker to the web page where the image was stored). We chose to use this design because a non-hyperlink design ensures that if a web page is invoked the picture is actually viewed, whereas in a hyperlink design it is possible for a participant to invoke the 20 web pages without actually viewing the images.

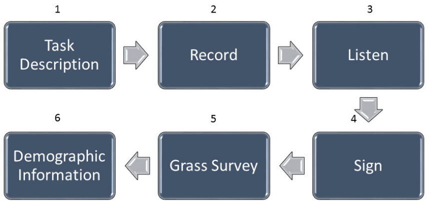


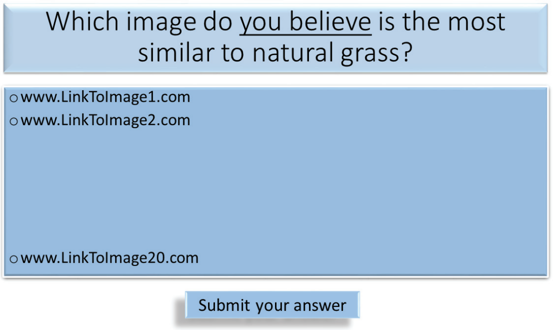
Figure 1: The experimental work-flow.

This task is a classic form of NGT task, as we ask the participants to state their *opinion*. Therefore it is a good fit to test our methods. In addition, this task does not require any special expertise or outstanding ability on the part of participants: the average person can complete this task exactly as described in the task directions.

The Methods

in this section we introduce the methods that were used in our experiments and explain how they were embedded in the experiment flow:

- Commitment by speech: In the second step of the experiment, participants were required to record themselves saying: "I promise to complete the tasks exactly the way stated in the task description." To ensure that participants recorded the above sentence, they were told that their work would be rejected if they failed to make the recording. In the third step of the experiment, participants were instructed to listen to a recording of a code and enter the code into a text box (similar to capcha). We indicated that this step was necessary to ensure they were not a robot. In the fourth step, participants were informed that they were about to begin the task and needed to check the box below and click the continue button to move forward.



Which image do you believe is the most similar to natural grass?

○ www.LinkToImage1.com
○ www.LinkToImage2.com
○ www.LinkToImage20.com

Figure 2: Illustration of the survey task.

- Commitment by listening: In the second step of the experiment, participants were required to record themselves saying: "hello world." Participants were told that their work would be rejected if they failed to record the sentence. We indicated that this step was necessary to ensure they were not a robot. In the third step of the experiment, participants were instructed to listen to a recording stating: "By clicking the agree button you promise to complete the tasks exactly the way stated in the task description." The recording ended with a code. Participants were required to enter it into a text box and then click a button stating "I agree" in order to continue to the next step. The fourth step was similar to the fourth step used in the commitment by speech experiment.
- Commitment by signature: The second step of the experiment was similar to the one used in the commitment by listening experiment (saying hello world). The third step was similar to the one used in the commitment by speech experiment (listening to a code). In the fourth part of the experiment, participants were presented with the following sentence: "I promise to complete the tasks exactly the way stated in the task description. To continue to the next step, participants were required to check a box and click a button stating "I agree."

We compared participants behavior in these conditions to two baseline experimental conditions:

- Baseline: The second step of the experiment was similar to the one used in the commitment by listening experiment (saying hello world). The third step of the experiment was similar to the one used in the commitment by speech experiment (listening to a code). The fourth step was similar to the fourth step used in the commitment by speech and commitment by listening experiments.
- Attention intervention: The second, third and fourth steps of the experiment were similar to those used in the baseline experiment. However, in this experiment, participants were not able to navigate between the fifth and the sixth steps of the experiment without spending at least three minutes on the actual task[1]. The attention intervention method is commonly used for drawing attention and awareness. The reason for using this intervention is to demonstrate that the lack of engagement exhibited by workers in our domain is not attributable to poor attention span.

Table 1 summarizes, for each experimental condition, the

[1] Participants were presented with a timer, and the submit button was only enabled when the timer reached zero.

	Record	Listen	Sign
C by speech	I promise to ...	code	continue to task
C by listening	hello world	I promise to ...	continue to task
C by signature	hello world	code	I promise to ...
Baseline	hello world	code	continue to task
Attention int.	hello world	code	continue to task

Table 1: Summary of each step of the experiment for the different conditions.

exact requirements in the record, listen, and sign interventions.

In several conditions, participants were required to record themselves saying hello world, listen to a recording of a code and check a box in order to continue to the sixth part of the experiment. The reason for requesting these inputs from the participants when they are obviously not a part of the treatments was to avoid participation bias. If these requirements were not included, the effect of the commitment methods could have been attributed to the people participating in each experiment (e.g., participants that are willing to record themselves are more engaged in the task). To avoid voice-related biases, the voice of the recordings at the third step of the experiment was generated by a computer (not human voice).

Platform & Participants

To test the proposed three methods, workers were recruited through Amazon Mechanical Turk (AMT).[2] We chose this platform because we attempted to emulate a crowd-sourcing scenario where an employer needs to recruit a large pool of workers for a NGT task. We did not apply any screening or qualification tests since the grass survey is not a complicated task that requires workers with special capabilities. Workers were paid 20 cents for their participation. To avoid a carryover effect, a between subject design was used. We recruited 5 groups of participants, each group had ~ 50 participants.[3] Most of the participants in the different experiment groups reported that they were located in the US and were 21-40 years old. As for gender, there was no significant difference[4].

Measures

We consider a successful intervention one that improves work quality. Measuring work quality can be achieved by measuring the output/work that was produced, analyzing workers behavior (e.g., task fingerprint) (Rzeszotarski and Kittur 2011), or both (Rzeszotarski and Kittur 2012). In NGT tasks, assessing the efficiency of an intervention is not straightforward. We measured other factors that have an indirect effect on work quality. We argue that the most important factor affecting work quality is worker engagement level. We use three measures to determine this factor in our experiments:

- Commitment: One of the most basic and fundamental indicators of workers engagement level is the way in which workers follow the job description. Hence, the first (and most important) measure we use is the binary measure of whether a worker completed the task exactly as stated in the task description. In our case, commitment means whether or not a worker viewed *all* the images before deciding which one to select. This measure by itself is not sufficient since there is no consideration for the quality of work; this measure only considers the fulfillment of the *minimal* task requirement. In our experiments, a worker who viewed all images but was not invested in the task (e.g., just viewing the images briefly, and picking an image randomly) would get a perfect score according to this measure.
- Dedication: This measures the *effort* a worker puts toward the task. In our case, it means the number of pictures a worker viewed before making a decision. While the first measure does not capture the difference between workers who are undedicated and workers who are semi-dedicated, the second measure differentiates between these groups of workers. Again, this measure by itself is not sufficient because viewing more images does not necessarily mean exerting more effort.
- Time: This measure is used to validate the results obtained by the other two measures. The time measure, in conjunction with the first two measures, presents a more holistic picture of the workers' engagement level. For each worker we calculate the average time s/he spent viewing one image: $\frac{\text{Task Completion Time}}{\text{Number of Images he/she Opened}}$. The underlying assumption is that more time a worker spent in viewing an image indicates better work quality. We use this measure to verify that quantity (i.e., number of images a worker views) does not come at the expense of quality (the time a worker invests in viewing the images). This measure by itself is not sufficient since it does not consider the amount of work. However the combination of the time put toward completion of the task and the amount of work done, indicates worker engagement in the task as well as the quality of the work itself.

Results

Figure 3 shows the effect of the different interventions over workers' commitment measure (the number of participants completing the task). In the baseline condition (no intervention), only 49% of the population viewed all 20 images before deciding. Most of the participants did not complete the task. As explained earlier, this percentage is an upper bound for the number of participants who were actually engaged in the task. In other words, out of the 49% that completed the task, we cannot tell how many participants were actually engaged in the task. This number could be much smaller. When the commitment by speech and commitment by listening interventions were used 70% of the population completed the task. We used nominal logistic regression with the commitment value (viewed all/some of the images) as the dependent variable, the different methods (Baseline, Attention, Signature, Listening, speech) as

[2] For a comparison between AMT and other recruitment methods see (Paolacci *et al.* 2010).

[3] We removed at most 3 participants from each group due to their failure to record themselves.

[4] For complete demographic information see Table 2.

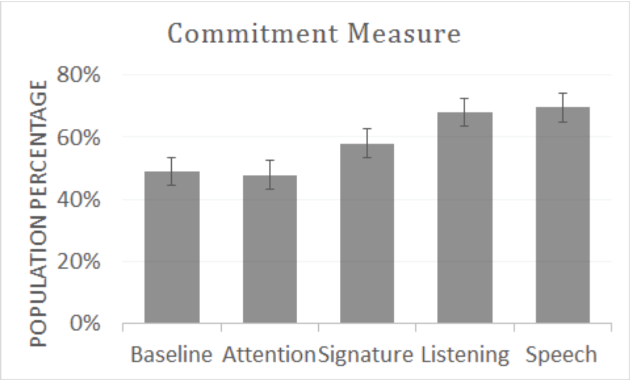


Figure 3: Percentage of population that completed the task in each one of the interventions.

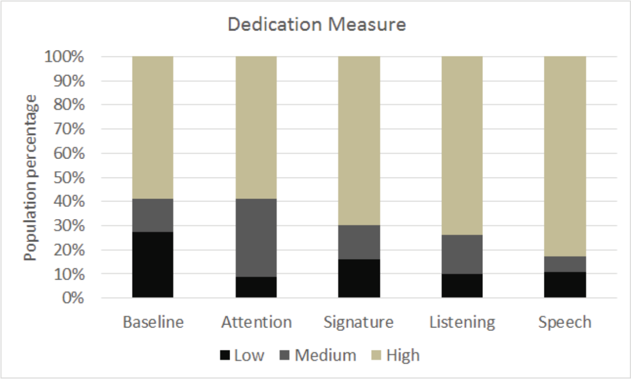


Figure 4: Percentage of population and the quantity of images they viewed in each one of the interventions.

factor, and user information (gender, age, location) as covariates. After controlling for the gender, location and age as covariates in a nominal logistics regeneration, the commitment by speech (M=0.7, SD=0.46) and commitment by listening (M=0.68, SD=0.47) interventions were found to improve significantly participants' commitment when comparing to the baseline condition. Commitment by speech yielded chiSq(df=1)=4.07,p=0.04, and commitment by listening yielded chiSq(df=1)=5.57,p=0.018. As for the commitment by signature intervention, while there is an improvement in the commitment measure when compared to the baseline, this improvement is not statistically significant chiSq(df=1)=0.04,p=0.83. Another interesting result shown in the figure is that the attention intervention did not improve commitment. This result is very surprising since this method provides a constant reminder to workers of what is expected from them, nevertheless their behavior did not change at all.

Figure 4 shows the extent to which workers completed the task (the dedication measure). Each bar in the graph is divided to three parts, the bottommost part represents the percentage of the population that made their decision after viewing less than 25% of the images (denoted 'Low'). The middle and topmost parts represent the percentage of the population that made their decision after $25\% - 75\%$, and *at least* 75% of the images (denoted 'Medium' and 'High' respectively). This graph complements Figure 3, and presents in more detail the amount of effort invested by participants for the different interventions. According to the commitment measure in Figure 3, the effect of commitment by listening and commitment by speech interventions was similar. However, Figure 4 reveals that the effect is not identical. When the commitment by listening intervention was applied 74% of the population were highly dedicated to the task, and when the commitment by speech intervention was applied 83% of the population were highly dedicated. This result indicates that there is a minor advantage for the speech intervention over the listening intervention. In addition, the dedication measure reveals that participants in the attention intervention condition behaved differently from participants in the baseline condition. This difference is not noticeable when considering just the commitment measure. The difference is not negligible, and it indicates that the attention intervention may affect peoples' behavior. In the baseline condition 27% of the participants exhibited low dedication, while in the attention intervention condition the percentage was significantly lower (9%). In other words, the attention intervention demonstrated ability to "push" a group of participants who exhibited low task engagement level to a stage were they are a little bit more engaged in the task.

We used logistic regression with the dedication value (number of images viewed) as the dependent variable, the different methods (Baseline, Attention, Signature, Listening, speech) as factor, and user information (gender,age,location) as covariates. After controlling for gender, location and age as covariates in logistics regeneration, the commitment by speech (M=17.1, SD=5.85) and commitment by listening (M=17.1, SD=5.4) interventions were found to improve significantly participants' commitment when comparing to the baseline condition. commitment by speech yielded F(1,88)=4.43,p=0.03 and commitment by listening yielded F(1,92)=5.43,p=0.02. As for the commitment by signature intervention (M=15.8,SD=6.4), while there is an improvement in the dedication measure when compared to the baseline (M=13.2, SD=8.2), this improvement is not statistically significant F(1,92)=0.97,p=0.3.

Figure 5 presents the average time a worker invested in examining one image (the time measure). In the figure, for each experimental condition, the population of participants is divided to three categories of dedication (similar to the categories presented in Figure 4). As can be seen in the figure, for 4 out of 5 conditions the average per image viewing time invested by participants who exhibited low, medium, and high dedication levels (different amount of images viewed), is similar. Since the average time participants invested in viewing images had nothing to do with the number of images they viewed we are able to verify that quantity does not come at the expense of quality, and therefore the number of images participants view is a good measure for work engagement.

The only discrepancy is in the attention intervention con-

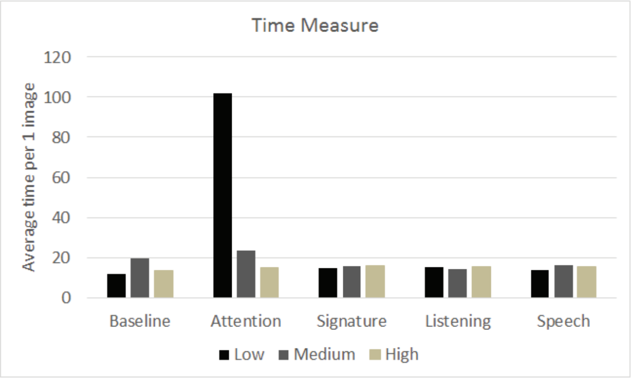


Figure 5: Average time (in seconds) participants invested watching each image.

dition. This discrepancy can be attributed to the nature of the intervention, which is based on time. A possible explanation could be that participants in this intervention viewed only a small portion of the images and then stared at the screen and waited for 3 minutes to pass so they would be able to click the “end-task” button.

Conclusions and Future Work

In this paper, we demonstrated the effectiveness of making a promise as a method for increasing worker engagement. The methods we presented were inspired by “commitment and consistency” theory. The innovation of this paper is in taking this psychological effect and harnessing it toward the improvement of work quality in crowdsourcing platforms. In addition, to the best of our knowledge there is no previous literature discussing what form of commitment is most effective. Through empirical investigation, we explored different forms of making commitment. Our results indicate that the most effective forms were commitment by speech and commitment by listening. This result is surprising because in traditional markets there is almost no value for an unwritten promise, while there is a high value for a promise made by a signature. The results presented in this work indicate that there is a higher value (in terms of commitment) for promises made by voice or by listening to a recording. We cannot determine why commitment by speech and listening were better than commitment by signature, but we hypothesize that the reason is the over use of signatures online. We hypothesize that people are used to this kind of interaction online and therefore do not consider it as a commitment mechanism. To further investigate this hypothesis more experiments are required.

In addition to presenting methods for improving worker engagement, we consider what factors made our methods work. The results from the attention intervention condition experiment assist us in concluding that the reason is not a result of attention related factors. Therefore, we hypothesize that the commitment action is the reason. Further experiments are required to prove this hypothesis.

This paper is a continuation of line of work dedicated to crowdwork quality improvement. The approach taken in this paper can be applied to any task without affecting worker preferences. This latter attribute is crucial to non-ground truth tasks, where usually the employee attempts to capture truthful preferences.

The approach taken in this work does not add a heavy burden for workers. The workers are not required to perform a time consuming or high cognitive load task. These two features are crucial to the success or failure of mechanism adoption. A good example for the importance of these two features can be the captcha mechanism for differentiating between human and robots. This mechanism is widely adopted by companies due to the minimal overhead and cognitive load it poses over the human users and because it is possible to apply the mechanism in any given platform.

The mechanism we offer in this work requires speakers or microphone. These two physical devices are very common and accessible to online labor workers. We note that it might be more difficult to recruit workers willing to record themselves since it requires workers to submit personal input which for some workers might be uncomfortable. In any case, since there is no statistical difference between the commitment by listening and commitment by speech interventions, we recommend using the commitment by listening mechanism.

While this paper offers a mechanism for NGT tasks, our findings may be relevant to GT tasks. Although there are already good mechanisms for motivating crowdworkers in GT task, combining our method with the existing mechanisms can further improve workers engagement. An obvious continuation work would be to compare the efficiency of the mechanisms presented in this paper with other mechanisms that were proven to be efficient in ground truth tasks.

There are two limitations of this work. First, to generalize our results to different domains, additional experimentations are required. Second, we cannot predict how our approach will affect people in the long run. It is possible that if our approach were to be adopted and widely used, crowd workers would accommodate promising and that would affect their commitment (i.e., the phenomenon exhibited in our experiment may “fade” over time if workers are constantly asked to promise.)

Appendix

		Baseline	Signature	Listen	Speech	Attention
Gender	Male	62%	44%	60%	46%	50%
	Female	38%	56%	40%	54%	50%
Location	US	60%	70%	64%	57%	61%
	India	12%	6%	16%	15%	28%
	Other	28%	24%	20%	28%	11%
Age	0-20	2%	8%	6%	4%	9%
	21-30	63%	46%	58%	43%	50%
	31-40	29%	26%	20%	30%	33%
	41-50	4%	12%	8%	15%	2%
	51- above	2%	8%	8%	7%	7%

Table 2: Demographic information of the participants in our experiments.

Acknowledgements

The research presented in this paper was partially supported by the Harvard Center for Research on Computation and Society.

References

Cem Akkaya, Alexander Conrad, Janyce Wiebe, and Rada Mihalcea. Amazon mechanical turk for subjectivity word sense disambiguation. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 195–203. Association for Computational Linguistics, 2010.

Michael S Bernstein, Greg Little, Robert C Miller, Björn Hartmann, Mark S Ackerman, David R Karger, David Crowell, and Katrina Panovich. Soylent: a word processor with a crowd inside. *Communications of the ACM*, 58(8):85–94, 2015.

Lukas Biewald. Massive multiplayer human computation for fun, money, and survival. In *International Conference on Web Engineering*, pages 171–176. Springer, 2011.

Jeffrey P Bigham, Chandrika Jayant, Hanjie Ji, Greg Little, Andrew Miller, Robert C Miller, Robin Miller, Aubrey Tatarowicz, Brandyn White, Samual White, et al. Vizwiz: nearly real-time answers to visual questions. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*, pages 333–342. ACM, 2010.

KD Borne and Zooniverse Team. The zooniverse: A framework for knowledge discovery from citizen science data. In *AGU Fall Meeting Abstracts*, volume 1, page 0650, 2011.

Robert B Cialdini and Nathalie Garde. *Influence*, volume 3. A. Michel, 1987.

Jackie Coyle-Shapiro and Ian Kessler. Consequences of the psychological contract for the employment relationship: A large scale survey. *Journal of management studies*, 37(7):903–930, 2000.

Anirban Dasgupta and Arpita Ghosh. Crowdsourced judgement elicitation with endogenous proficiency. In *Proceedings of the 22nd international conference on World Wide Web*, pages 319–330. ACM, 2013.

Ryan Drapeau, Lydia B Chilton, Jonathan Bragg, and Daniel S Weld. Microtalk: Using argumentation to improve crowdsourcing accuracy. In *Proceedings of the 4th AAAI Conference on Human Computation and Crowdsourcing (HCOMP)*, 2016.

Zhenni Feng, Yanmin Zhu, Qian Zhang, Lionel M Ni, and Athanasios V Vasilakos. Trac: Truthful auction for location-aware collaborative sensing in mobile crowdsourcing. In *Proceedings of INFOCOM*, pages 1231–1239, 2014.

Xi Alice Gao, Yoram Bachrach, Peter Key, and Thore Graepel. Quality expectation-variance tradeoffs in crowdsourcing contests. In *Proc. of AAAI*, 2012.

Anthony G Greenwald, Catherine G Carnot, Rebecca Beach, and Barbara Young. Increasing voting behavior by asking people if they expect to vote. *Journal of Applied Psychology*, 72(2):315, 1987.

Chien-Ju Ho, Aleksandrs Slivkins, Siddharth Suri, and Jennifer Wortman Vaughan. Incentivizing high quality crowdwork. In *Proc. of WWW*, pages 419–429, 2015.

Tobias Hossfeld, Christian Keimel, Matthias Hirth, Bruno Gardlo, Julian Habigt, Klaus Diepold, and Phuoc Tran-Gia. Best practices for qoe crowdtesting: Qoe assessment with crowdsourcing. *IEEE Transactions on Multimedia*, 16(2):541–558, 2014.

Gary Hsieh and Rafał Kocielnik. You get who you pay for: The impact of incentives on participation bias. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, pages 823–835. ACM, 2016.

Shih-Wen Huang and Wai-Tat Fu. Systematic analysis of output agreement games: Effects of gaming environment, social interaction, and feedback. *Urbana*, 51:61801, 2012.

Shih-Wen Huang and Wai-Tat Fu. Don't hide in the crowd!: increasing social transparency between peer workers improves crowdsourcing outcomes. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 621–630. ACM, 2013.

Shih-Wen Huang and Wai-Tat Fu. Enhancing reliability using peer consistency evaluation in human computation. In *Proceedings of the 2013 conference on Computer supported cooperative work*, pages 639–648. ACM, 2013.

D. Kahneman. A psychological point of view: Violations of rational rules as a diagnostic of mental processes (commentary on stanovich and west). *Behavioral and Brain Sciences*, 23:681–683, 2000.

Ece Kamar and Eric Horvitz. Incentives for truthful reporting in crowdsourcing. In *Proceedings of the 11th international conference on autonomous agents and multiagent systems-volume 3*, pages 1329–1330. International Foundation for Autonomous Agents and Multiagent Systems, 2012.

Nicolas Kaufmann, Thimo Schulze, and Daniel Veit. More than fun and money. worker motivation in crowdsourcing-a study on mechanical turk. In *AMCIS*, volume 11, pages 1–11, 2011.

Juho Kim, Phu Tran Nguyen, Sarah Weir, Philip J Guo, Robert C Miller, and Krzysztof Z Gajos. Crowdsourcing step-by-step information extraction to enhance existing how-to videos. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems*, pages 4017–4026. ACM, 2014.

Aniket Kittur, Ed H Chi, and Bongwon Suh. Crowdsourcing user studies with mechanical turk. In *Proc. of CHI*, pages 453–456, 2008.

Aniket Kittur, Boris Smus, Susheel Khamkar, and Robert E Kraut. Crowdforge: Crowdsourcing complex work. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*, pages 43–52. ACM, 2011.

Aniket Kittur, Susheel Khamkar, Paul André, and Robert Kraut. Crowdweaver: visually managing complex crowd work. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, pages 1033–1036. ACM, 2012.

Aniket Kittur, Jeffrey V Nickerson, Michael Bernstein, Elizabeth Gerber, Aaron Shaw, John Zimmerman, Matt Lease, and John Horton. The future of crowd work. In *Proceedings of the 2013 conference on Computer supported cooperative work*, pages 1301–1318. ACM, 2013.

Anand P Kulkarni, Matthew Can, and Bjoern Hartmann. Turkomatic: automatic recursive task and workflow design for mechanical turk. In *CHI'11 Extended Abstracts on Human Factors in Computing Systems*, pages 2053–2058. ACM, 2011.

Martha Larson, Mark Melenhorst, María Menéndez, and Peng Xu. Using crowdsourcing to capture complexity in human interpretations of multimedia content. In *Fusion in Computer Vision*, pages 229–269. Springer, 2014.

Walter S Lasecki, Christopher D Miller, and Jeffrey P Bigham. Warping time for more effective real-time crowdsourcing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2033–2036. ACM, 2013.

Edith Law, Ming Yin, Joslin Goh, Kevin Chen, Michael A Terry, and Krzysztof Z Gajos. Curiosity killed the cat, but makes crowdwork better. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 4098–4110. ACM, 2016.

Florian Laws, Christian Scheible, and Hinrich Schütze. Active learning with amazon mechanical turk. In *Proceedings of the conference on empirical methods in natural language processing*, pages 1546–1556. Association for Computational Linguistics, 2011.

Greg Little, Lydia B Chilton, Max Goldman, and Robert C Miller. Turkit: human computation algorithms on mechanical turk. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*, pages 57–66. ACM, 2010.

Yang Liu and Yiling Chen. Learning to incentivize: Eliciting effort via output agreement. In *In the Proc. of the 25th International Joint Conference on Artificial Intelligence (IJCAI'16), 2016*, July 2016.

Babak Loni, Lei Yen Cheung, Michael Riegler, Alessandro Bozzon, Luke Gottlieb, and Martha Larson. Fashion 10000: an enriched social image dataset for fashion and clothing. In *Proceedings of the 5th ACM Multimedia Systems Conference*, pages 41–46. ACM, 2014.

Andrew Mao, Ece Kamar, Yiling Chen, Eric Horvitz, Megan E Schwamb, Chris J Lintott, and Arfon M Smith. Volunteering versus work for pay: Incentives and tradeoffs in crowdsourcing. In *Proc. of HCOMP*, 2013.

Winter Mason and Duncan J Watts. Financial incentives and the performance of crowds. *ACM SigKDD Explorations Newsletter*, 11(2):100–108, 2010.

G. Paolacci, J. Chandler, and P. Ipeirotis. Running experiments on amazon mechanical turk. *Judgment and Decision Making*, 5(5):411–419, 2010.

Genevieve Patterson and James Hays. Sun attribute database: Discovering, annotating, and recognizing scene attributes. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2751–2758. IEEE, 2012.

Genevieve Patterson, Chen Xu, Hang Su, and James Hays. The sun attribute database: Beyond categories for deeper scene understanding. *International Journal of Computer Vision*, 108(1-2):59–81, 2014.

Dražen Prelec. A bayesian truth serum for subjective data. *science*, 306(5695):462–466, 2004.

Goran Radanovic and Boi Faltings. Learning to scale payments in crowdsourcing with properboost. In *Proceedings of the Fourth AAAI Conference on Human Computation and Crowdsourcing (HCOMP" 16)*, number EPFL-CONF-222809, 2016.

Vikas C Raykar, Shipeng Yu, Linda H Zhao, Gerardo Hermosillo Valadez, Charles Florin, Luca Bogoni, and Linda Moy. Learning from crowds. *Journal of Machine Learning Research*, 11(Apr):1297–1322, 2010.

Katharina Reinecke and Krzysztof Z. Gajos. Labinthewild: Conducting large-scale online experiments with uncompensated samples. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work &#38; Social Computing*, CSCW '15, pages 1364–1378, New York, NY, USA, 2015. ACM.

Sandra L Robinson. Trust and breach of the psychological contract. *Administrative science quarterly*, pages 574–599, 1996.

Jeffrey M Rzeszotarski and Aniket Kittur. Instrumenting the crowd: using implicit behavioral measures to predict task performance. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*, pages 13–22. ACM, 2011.

Jeffrey Rzeszotarski and Aniket Kittur. Crowdscape: interactively visualizing user behavior and output. In *Proceedings of the 25th annual ACM symposium on User interface software and technology*, pages 55–62. ACM, 2012.

Janet Schwartz, Daniel Mochon, Lauren Wyper, Josiase Maroba, Deepak Patel, and Dan Ariely. Healthier by precommitment. *Psychological science*, 25(2):538–546, 2014.

Steven J Sherman. On the self-erasing nature of errors of prediction. *Journal of personality and Social Psychology*, 39(2):211, 1980.

Jiwoong Shin and Dan Ariely. Keeping doors open: The effect of unavailability on incentives to keep options viable. *Management Science*, 50(5):575–586, 2004.

Herbert Alexander Simon. *Models of bounded rationality: Empirically grounded economic reason*, volume 3. MIT press, 1982.

Adish Singla and Andreas Krause. Truthful incentives in crowdsourcing tasks using regret minimization mechanisms. In *Proceedings of the 22nd international conference on World Wide Web*, pages 1167–1178. ACM, 2013.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y Ng. Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of the conference on empirical methods in natural* *language processing*, pages 254–263. Association for Computational Linguistics, 2008.

Mohammad Soleymani, Maja Pantic, and Thierry Pun. Multimodal emotion recognition in response to videos. *IEEE transactions on affective computing*, 3(2):211–223, 2012.

Jens Witkowski and David C Parkes. Peer prediction without a common prior. In *Proceedings of the 13th ACM Conference on Electronic Commerce*, pages 964–981. ACM, 2012.

Ming Yin and Yiling Chen. Bonus or not? learn to reward in crowdsourcing. In *IJCAI*, pages 201–208, 2015.

Ming Yin, Yiling Chen, and Yu-An Sun. Monetary interventions in crowdsourcing task switching. In *Proc. of HCOMP*, 2014.