Motivating Quality Work

— W241 Final Project —

Kevin Hanna, Kevin Stone, Changjing Zhao

Background

According to Forbes, 2018 was the year of awakening and 2019 is the year of action for AI. At the end of 2019 over half of the companies surveyed have implemented some form of AI in their business. Before companies can implement and deploy their AI models to make predictions, humans often have to manually annotate thousands of examples and add the labels required to train the machine learning models. This process is time consuming and expensive. Given the growing demand for AI applications, the appetite for labeling datasets will continue unabated for the foreseeable future. Crowdsourced workers can be a cost-effective option to label datasets, but understandably ML practitioners want quality-minded crowdsourced workers sourced at a reasonable cost.

In terms of crowdsourcing, research in psychology and sociology provide a theoretical basis on human motivation. However, these approaches have to be adapted for a crowdsourcing setting. At the same, by motivating workers to contribute more, requesters can unwillingly make them more susceptible to quality control issues, so careful motivational considerations should be taken into account [1].

[1] proposes a model that classifies motivation into two types - intrinsic and extrinsic. Intrinsic motivation involves doing something because it's personally rewarding to you and can be broken down into enjoyment-based and community-based motivation. Extrinsic motivation involves doing something because you want to earn a reward or avoid punishment and can be broken down into immediate payoff, future payoff and social motivation. Examples of intrinsic and extrinsic motivations are shown in Table 1.

Intrinsic Motivation	Enjoyment-based	a worker does a translation task because he likes translating and wants to use his skills in his favorite foreign language or the task allows him to be creative (e.g. designing a logo or a website)
	Community-based	a worker does a task that allows her to be active on a crowdsourcing platform to meet new people.
Extrinsic Motivation	Immediate payoff	a worker does a task as a form of income
	Future payoff	a worker does a translation tasks because she wants to improve language skills for a new or better job
	Community-based	a worker does a task because he seeks commendation

Table 1. Examples of intrinsic and extrinsic motivations

Research Question

The scope of our research is to understand what motivates crowdsourced workers to do quality work. Among understanding whether intrinsic and extrinsic motivation drives quality work, we focus on the three sub-categories of extrinsic motivation: immediate payoff, future payoff, and community-based.

Hypothesis

Our hypotheses based on our research question are listed in Table 2.

Immediate payoff	quality will increase with a higher financial reward quality will increase with negative financial reward
Future payoff	quality will increase with motivation that benefits the future
Community-based	quality will increase with explicit request for quality (e.g. please do good work) quality will reduce with negative motivation (e.g. helping develop a government surveillance app)

Table 2. Hypotheses for each category of extrinsic motivation

Experiments

Experimental Setup

We decided to use Mechanical Turk workers (MTurkers) as our experimental subjects to measure the quality of work based on different motivations. The MTurkers are randomly selected to receive either the treatment or control. The randomization process is described in more detail below. Whether in treatment or control, each MTurker is asked to draw a bounding box around a car. The car images were selected from the Open Images Dataset V5. As part of our pilot test, we asked multiple MTurkers to draw a bounding box around 20 cars, one car per image. A sample of four of these images is shown in Figure 1 below.

Figure 1. Sample of 20 images used to select final image for experiment

The quality of the MTurkers' work is assessed by measuring how far their bounding box is from the ground truth. This accuracy measure is calculated by summing the Euclidean distances between the top-left and bottom-right corners of the MTurkers' bounding box and the ground truth bounding box. A visual of the accuracy calculation can be seen in Figure 2.

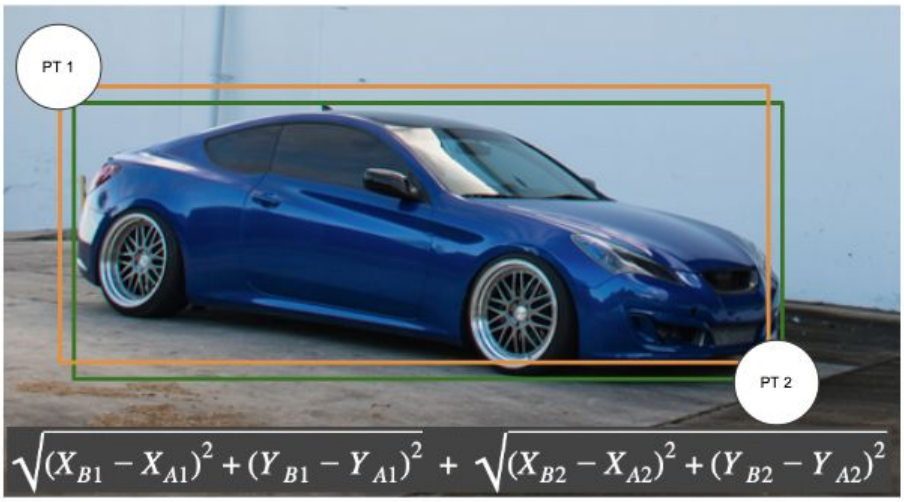


Figure 2. Accuracy calculation. Green box is ground truth and orange box is the MTurker's bounding box.

We chose an image that had a medium accuracy score compared to the other 19 images, which is the image of the blue car in the bottom right in Figure 1 and in Figure 2. We then placed the image in the Mechanical Turk bounding box tool as shown in Figure 4. Actual treatment and control instructions replaced the "Treatment/Control Text Here" text. Prior to seeing the bounding box tool MTurkers in control and treatment were shown the instructions in Figure 3.

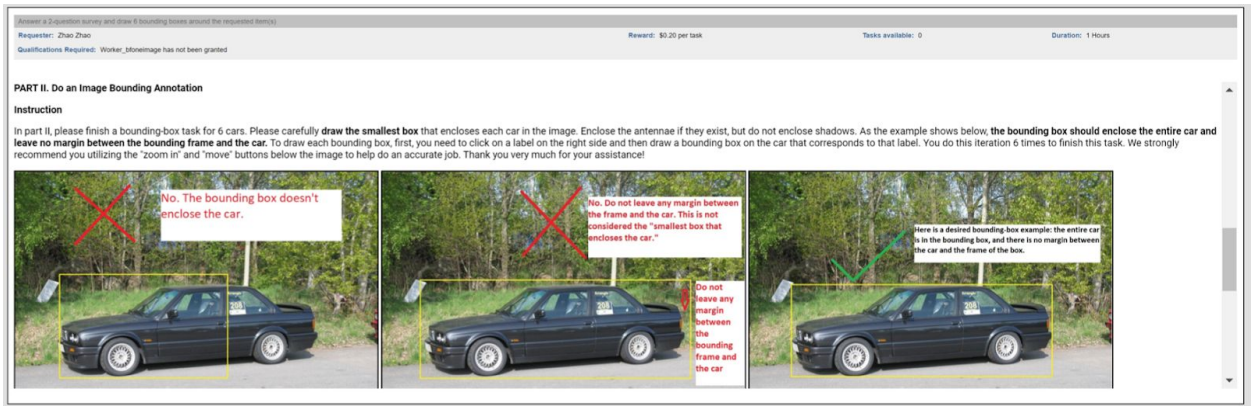


Figure 3. Initial instructions for MTurkers for both treatment and control.

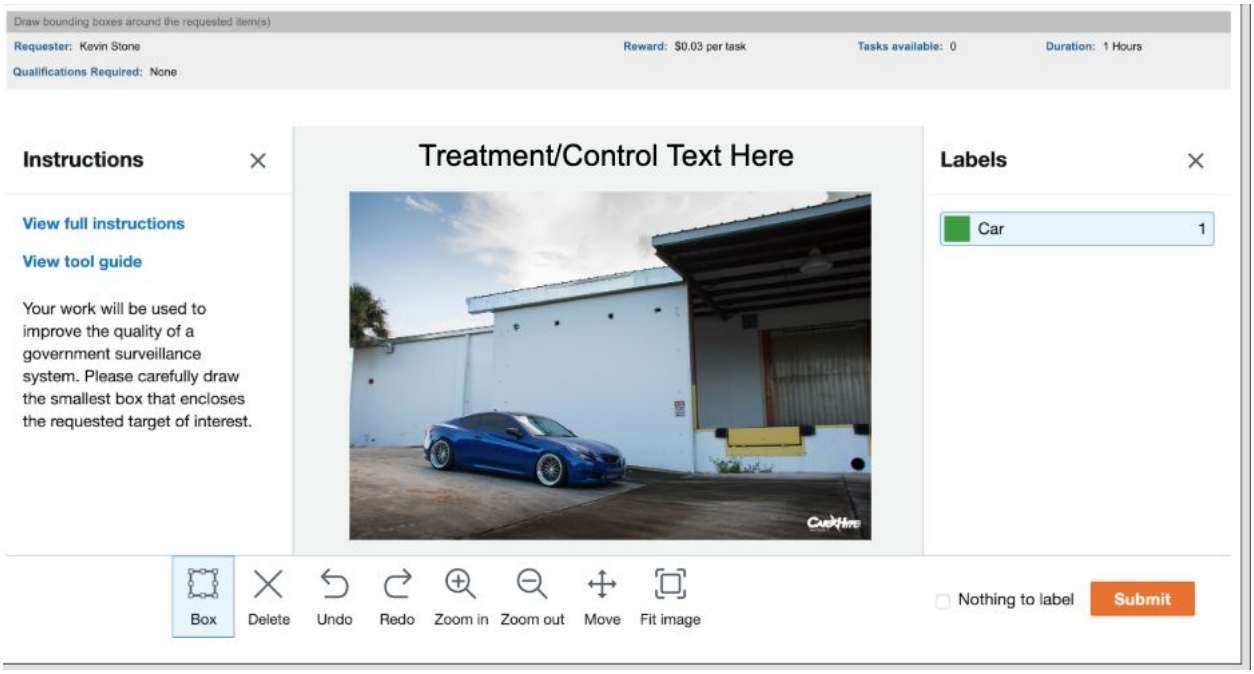


Figure 4. Bounding box tool used for both treatment and control.

Randomization

As part of our pilot testing, we collected survey data (see Appendix for questions). We found that less than 50% of MTurkers who completed our survey responded to our subsequent request to participate in a bounding box experiment. Because of this extremely high attrition rate, we were unable to randomize from a set of known MTurkers. We moved to using convenience sampling as our pseudo-randomization methodology. For each experiment, we collected data in a few batches, in each of which, we collect a percentage of control and treatment. Additionally, we also vary the sequence of collecting treatment and control from batch to batch. An example of this approach is shown in Figure 5.

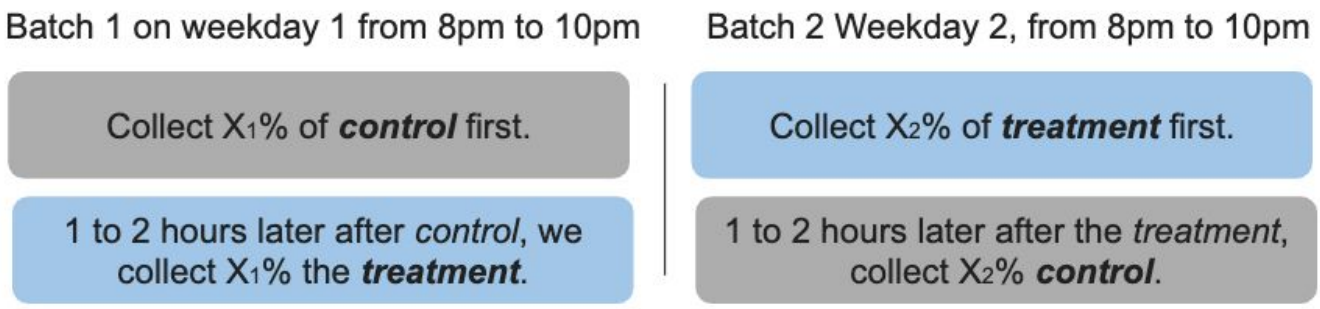


Figure 5. Example of technique to improve randomization of convenience sampling

Experimental Design

We use a posttest treatment design as can be seen in the chart below:

Treatment Group	R	X	O
Control Group	R	—	O

Pilot Overview

For our pilot, we gave the MTurkers a negative treatment and asked that they draw a single bounding box on each of 20 images. We first collected some information about the subject through a survey and then randomly assigned those subjects to treatment and control. Our primary goal was to understand how our scoring scheme worked, gauge the level of variance we should expect in future experiments and test if our covariates collected from our survey were helpful. We had high attrition and due to a misunderstanding of the Mechanical Turk platform, our assignments to treatment and control failed and we ended up with MTurkers in both treatment and control and others not in our results at all.

We were not able to produce a valid ATE, but we could at least see the variance, which was high.

Experiment Overview

As summary of the experiments is shown in Table 3 below. Although the accuracy of Mturker work is important their tasks often do not require a high-level of skills to perform. As such they are often low-paid tasks. Humans tend to try to do a good job if they are aware that their work matters. Therefore, in Experiment I we decided to include the messaging “Your work will be used to develop a self-driving car system.” to motivate the workers. We also tried a negative messages for this category, “Your work will be used to develop a government surveillance system,” hoping to see the accuracy being worse than the control.

Group/Motivation	Experiment I: Social Motivation	Experiment II: Future Payoff	Experiment III: Immediate Payoff	Experiment IV: All Three Motivations Combined
Control	Pay 2 cents + **No messaging**	Pay 20 cents + **No messaging**	Pay 5 cents + **No messaging**	Pay 20 cents + **No messaging**
Treatment	Pay 2 cents + Treatment: **"Your work will be used to develop a government surveillance system."**	Pay 20 cents + Treatment: **"We use the quality of this bounding-box task to help decide if we would assign our future tasks to you for assistance"**	Pay 5 cents + Treatment: **"Pay 5 cents only if job meets insturctions."**	Pay 40 cents + **All three motivations combined**
Task and Per-Group Sample Size	Do a bounding box; N = 200	Answer a survey and do a bounding box; N = 350	Do a bounding box; N = 200	Do 6 bounding boxes; N = 400

Table 3. Summary of Experiments

Experiment II's messaging was inspired by feedback from the workers. A few workers emailed us their feedback about our tasks, expressing their willingness to do more similar tasks for us. Maintaining a good reputation to acquire stable task opportunities fell in the "future payoff" category, and we then decided to try this messaging, "We use the quality of this bounding-box task to help decide if we would assign our future tasks to you for assistance."

In Experiment III we varied both payment amount and payment messages. According to current studies [1], the primary motivation of Mturker is for immediate rewards. By raising the payment amount and stating a warning about declining their work, we hoped to see that bounding boxes in the treatment group were more accurate than those in the control group, in which workers were paid less than half of the amount and were not given the warning message.

Lastly, in Experiment IV, we threw in the who kitchen sink (all three motivations combined) to see if we could produce an effect.

Data Completeness

Attrition is no longer a big issue when we collect data within one task, and what attrition does exist is surprisingly well balanced between treatment and control in each experiment. Note that attrition in our experiments is defined as not completing a bounding box around the image.

experiment	in_treatment	Total	Attriters
future	1	350	6
future	0	349	6
immediate	0	200	1
immediate	1	200	1
pilot	0	400	0
pilot	1	400	1
social	0	200	1
social	1	200	2

Table 4. Summary of attriters for each experiment

As Table 4 shows, the number of attriters is low compared to the totals and is evenly distributed between treatment and control. We have therefore decided to exclude the attriters from the analyses. If attrition were to be a concern, we would calculate a range of treatment effects by assuming the attriters of each group to have the minimum and maximum values of the accuracy score of each group respectively. From using both the minimum imputation and the maximum imputation, we would be able to report the range for the treatment effect.

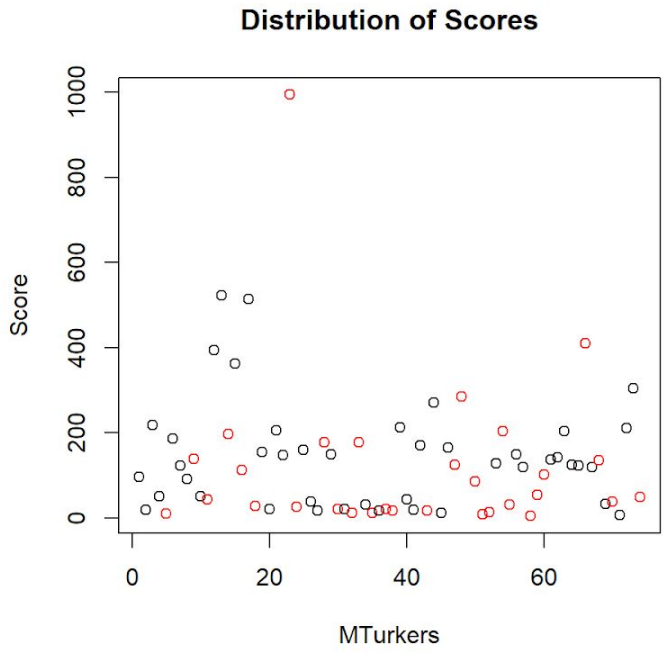
Analysis

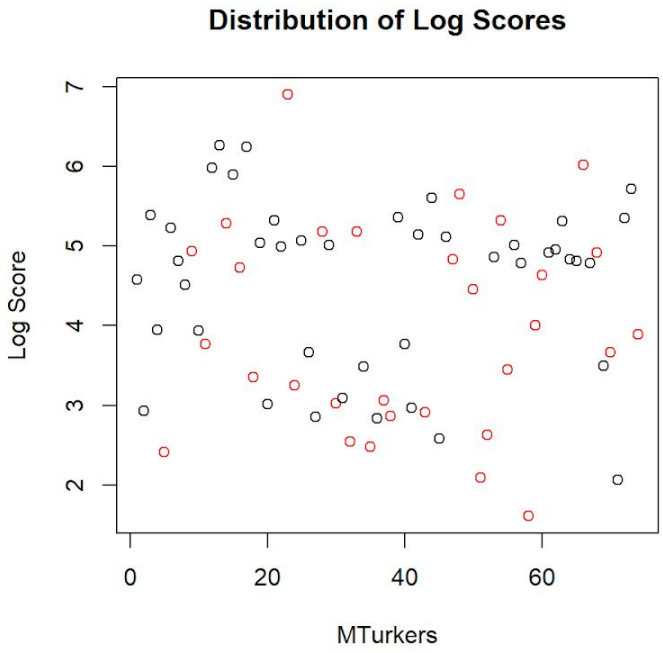
Pilot Study

EDA (N = 40 in Treatment or Control)

For our pilot, we gave the Turkers a social treatment suggesting their work would be used to improve a government surveillance system and asked that they draw a single bounding box on each of 20 images. We first collected some information about the subject through a survey and then randomly assigned those subjects to treatment and control. Our primary goal was to understand how our scoring scheme worked, gauge level of variance we should expect in future experiments and test if our covariates collected from our survey explained any of the variance. We had high attrition and due to a misunderstanding of the Mechanical Turk platform, our assignments to treatment and control failed and we ended up with Turkers not in our experiment in our results, and many ended up in both treatment and control.

We were not able to trust any ATE, but we could at least see the variance, which was exceptionally high. In either chart below, the X axis represented the i-th MTurker worker, and the worker sequence was random. 80 workers were in this test. The Y axis represented the bounding box accuracy score. The majority of the accuracy score ranged from 0 to around 300. Red dots, spreading with a wider range, represented the treatment and black dots were the control.





○ Control ○ Treatment

in_treatment	mean_score	std_dev
0	148.0579	124.3664
1	115.7390	187.7335

mean_worker_score
Min. : 5.003
1st Qu.: 31.396
Median :116.653
Mean :134.519
3rd Qu.:176.462
Max. :994.601
NA's :1

Experiment I - Social Treatment

With the first pilot behind us, we decided we needed to focus on increasing our statistical power and hypothesized that collecting the same number of bounding boxes but using more subjects & fewer experiments would provide more statistical power. Each subject was presented with a single image and created a single bounding box.

The charts below showed the accuracy score and logged accuracy score of a random i-th worker. Red dots represented the treatment group and black dots represented the control. The charts didn't visually reveal that there was a clear treatment effect. We would verify it with the following analyses.

EDA (N = 200 in Treatment or Control)

in_treatment	mean_score	std_dev
0	15.60004	35.84162
1	19.55064	48.77030

bounding_box_score
Min. : 1.423
1st Qu.: 5.226
Median : 8.946
Mean : 17.570
3rd Qu.: 14.294
Max. :489.540
NA's :3

Regression Analysis

The results of our regression failed to show any reliable effect of our treatment. The coefficient of "in_treatment" is 3.951, indicating the message about government surveillance usage caused workers to produce a worse accuracy. But with a p-value of 0.36 there, no information can be gleaned from this with any confidence.

With this experiment, the other variable on which we could evaluate the treatment effect was the amount of time each Turker spent on the task. And working time doesn't seem to be affected by our treatment.

Treatment Effect on Bounding Score Accuracy

Treatment Messaging: "Your work will be used to develop a government surveillance system."

	Dependent variable:
	bounding_box_score
in_treatment	3.951
	p = 0.359
Constant	15.600***
	p = 0.00000
Observations	397
R^2	0.002
Adjusted R^2	−0.0004
Residual Std. Error	42.781 (df = 395)
F Statistic	0.846 (df = 1; 395)
Note:	*p<0.1; **p<0.05; ***p<0.01

Treatment Effect on Time Spent on Task

Treatment Messaging: "Your work will be used to develop a government surveillance system."

	Dependent variable:
	WorkTimeInSeconds
in_treatment	−12.060
	p = 0.475
Constant	89.920***
	p = 0.000
Observations	400
R^2	0.001
Adjusted R^2	−0.001
Residual Std. Error	168.619 (df = 398)
F Statistic	0.512 (df = 1; 398)
Note:	*p<0.1; **p<0.05; ***p<0.01

The results suggest the negative treatment caused Turkers to spend less time on the task, but the p-value is far from statistically significant again.

Additionally, we evaluated the relationship between time spent on task and the accuracy score. Because both variables were highly right skewed, we took a log transformation on both and calculated the correlation to be a low negative one, -0.17, indicating that to generate a more accurate result (lower score), workers needed slightly longer time.

Power Test

To achieve the statistical power of 0.8 at the 0.05 confidence-level with the variance we had in this experiment, we would require nearly 1450 subjects in both control and treatment.

```
Two-sample t test power calculation

         n = 1450.123
     delta = 3.950596
        sd = 42.77251
 sig.level = 0.05
     power = 0.8

  alternative = one.sided
```

Learnings from This Experiment

The estimated 2,900 subjects required to achieve the statistical power we needed was too many. With a p-value of 0.359, even with the 2,900 subjects, we weren't likely to find a statistically significant ATE. We need to change our experiment and collect more covariates.

Experiment II - Future Payoff

In Experiment I, we used a treatment which we hypothesized would cause the MTurkers in treatment to work less hard and as a result the ATE was positive, which in our scoring means the bounding was less accurate. We also wanted to test if a treatment promoting higher accuracy would have a more significant ATE, so the MTurkers in treatment were told we were looking for MTurkers to perform some future work with the hypothesis

that if the MTurkers thought of the task as a test with the incentive of future work they would try harder, and result in a better score. So we ran a small experiment to test this hypothesis.

Regression Analysis

At first look there doesn't seem to be any significant treatment affect, the p-value had gone down from 0.36 in the previous experiment to 0.28 in this, but we only used a quarter the number of subjects.

Treatment Messaging: "We use the quality of this bounding box task to help decide if we would assign our future tasks to you."

	Dependent variable:
	bounding_box_score Future Payoff
in_treatment	6.956 p = 0.281
Constant	18.654*** p = 0.0001
Observations	97
R^2	0.012
Adjusted R^2	0.002
Residual Std. Error	31.555 (df = 95)
F Statistic	1.177 (df = 1; 95)
Note:	*p<0.1; **p<0.05; ***p<0.01

Covariate Regression Analysis

In this experiment we asked the Turkers to answer some questions about the device they were using, their experience doing these types of tasks and some demographic information.

	Dependent variable:		
	bounding_box_score		
	Target Alone	Monitor size	Did task before
	(1)	(2)	(3)
in_treatment	6.956	10.541	7.274
	p = 0.281	p = 0.117	p = 0.275
monitorlargescreen		−65.612***	
		p = 0.001	
monitormidsize		−57.717***	
		p = 0.002	
monitorsmalllaptop		−56.840***	
		p = 0.003	
monitortablet		−33.229*	
		p = 0.095	
didbfno			11.372
			p = 0.471
didbfyes			7.336
			p = 0.619
Constant	18.654***	71.057***	9.539
	p = 0.0001	p = 0.0002	p = 0.492
Observations	97	97	95
R^2	0.012	0.179	0.027
Adjusted R^2	0.002	0.133	−0.005
Residual Std. Error	31.555 (df = 95)	29.402 (df = 91)	30.884 (df = 91)
F Statistic	1.177 (df = 1; 95)	3.955*** (df = 5; 91)	0.841 (df = 3; 91)
Note:			*p<0.1; **p<0.05; ***p<0.01

	Dependent variable:		
	bounding_box_score		
	Education	Income	Age
	(1)	(2)	(3)
in_treatment	6.284	2.807	5.999
	p = 0.338	p = 0.691	p = 0.311
eduhighschool	−17.107		
	p = 0.453		
edumasterorabove	−15.088		
	p = 0.127		
edusomecollege	−17.204		
	p = 0.171		
incomegt30klt60k		10.648	
		p = 0.203	
incomegt60klt90k		−4.758	
		p = 0.650	
incomegt90k		17.478	
		p = 0.209	
incomelt10k		−9.106	
		p = 0.409	
age31to40			−10.309
			p = 0.365
age41to50			96.295***
			p = 0.00001
agelto21			−12.077
			p = 0.678
Constant	22.440***	17.499***	18.000***
	p = 0.00002	p = 0.003	p = 0.0001
Observations	97	97	97
R^2	0.056	0.072	0.214
Adjusted R^2	0.015	0.021	0.180
Residual Std. Error	31.342 (df = 92)	31.250 (df = 91)	28.609 (df = 92)
F Statistic	1.371 (df = 4; 92)	1.412 (df = 5; 91)	6.251*** (df = 4; 92)
Note:			*p<0.1; **p<0.05; ***p<0.01

All of the coefficients for the screen size question were negative, and by a fairly significant amount. The baseline value was cellphone, which can be significantly smaller than all the other types of screens. So we tested that on its own.

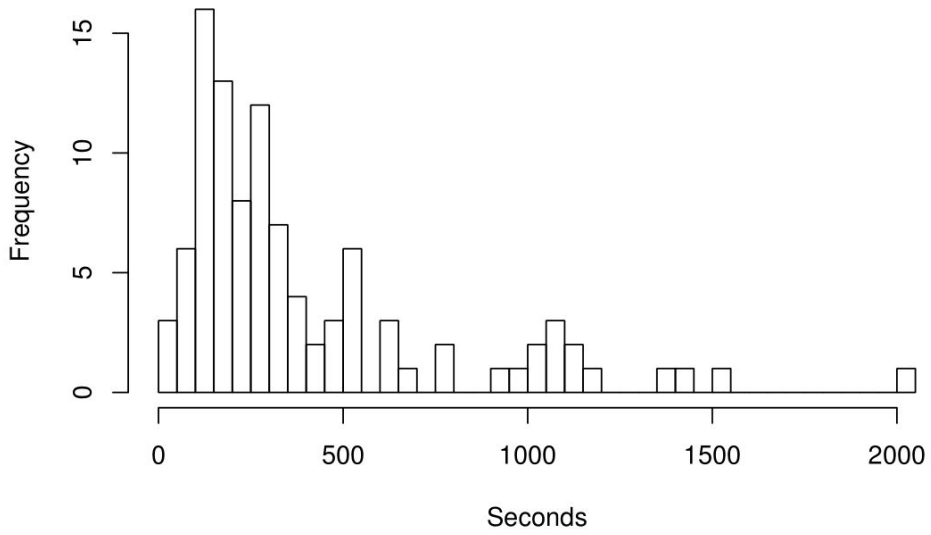
	Dependent variable:	
	bounding_box_score	
	Target Alone	Mobile
	(1)	(2)
in_treatment	6.956	14.153**
	p = 0.281	p = 0.029
is_mobile		34.032***
		p = 0.0003
Constant	18.654***	10.400**
	p = 0.0001	p = 0.033
Observations	97	95
R^2	0.012	0.146
Adjusted R^2	0.002	0.128
Residual Std. Error	31.555 (df = 95)	29.759 (df = 92)
F Statistic	1.177 (df = 1; 95)	7.890*** (df = 2; 92)
Note:		*p<0.1; **p<0.05; ***p<0.01

If the subject is using a cellphone to do the task, their accuracy goes down (score increases), which is intuitive. Having a cellphone as a control decreases the p-value from 0.28 to a fairly significant 0.029. This result is unexpected in that the treatment seems to be causing the opposite of the hypothesized effect.

As with the previous experiment, we also analyzed how the treatment affected the amount of time they spent on the task.

	Dependent variable:	
	WorkTimeInSeconds	
	Future Payoff	Social
	(1)	(2)
in_treatment	12.880	−12.060
	p = 0.866	p = 0.475
Constant	394.260***	89.920***
	p = 0.000	p = 0.000
Observations	100	400
R^2	0.0003	0.001
Adjusted R^2	−0.010	−0.001
Residual Std. Error	379.849 (df = 98)	168.619 (df = 398)
F Statistic	0.029 (df = 1; 98)	0.512 (df = 1; 398)
Note:		*p<0.1; **p<0.05; ***p<0.01

The regression shows that those in our future payoff treatment on average spent 13 seconds more time, the opposite from our previous treatment, which is what we hypothesize, however, the p-values is quite large.



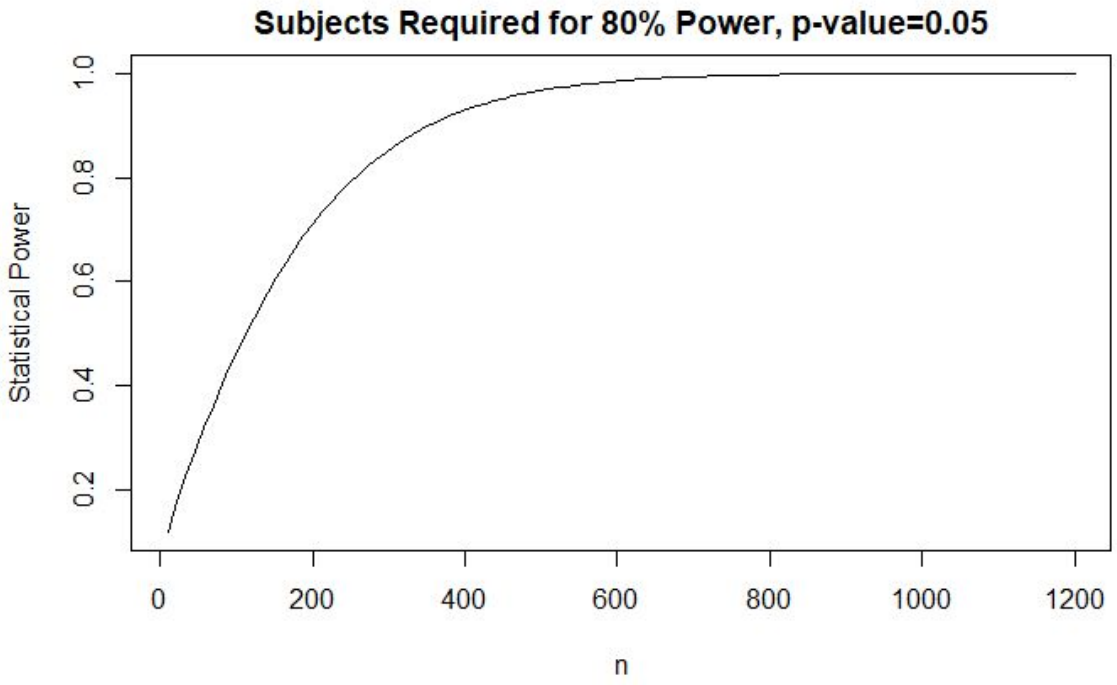
Many subjects are spending well over the reasonable time it should take to perform this task of answering 6 multiple-choice questions about themselves and drawing a single bounding box around a car. suggesting that Turkers might not be concentrating on our task alone, it could be they are spawning multiple tabs. Regardless, working time is not helpful for our experiment.

Power Test

The power calculation when using the negative treatment, telling those in treatment that they were doing work for a government surveillance system estimated we needed 2,900 subjects. Using an incentive of possible future work as the treatment, the ATE has less variance, and estimated that we only need 255 subjects in each group to get 0.80 statistical power, so well worth gathering more data.

```
Two-sample t test power calculation

              n = 255.6101
          delta = 6.955859
             sd = 31.5837
      sig.level = 0.05
          power = 0.8
    alternative = one.sided
```



Additional data collected

To improve the statistical power for this experiment, we collected data from 600 more subjects.

	Dependent variable:	
	bounding_box_score	
	n=700	n=100
	(1)	(2)
in_treatment	−13.050**	6.956
	p = 0.032	p = 0.281
Constant	28.934***	18.654***
	p = 0.000	p = 0.0001
Observations	687	97
R^2	0.007	0.012
Adjusted R^2	0.005	0.002
Residual Std. Error	79.528 (df = 685)	31.555 (df = 95)
F Statistic	4.625** (df = 1; 685)	1.177 (df = 1; 95)
Note:		*p<0.1; **p<0.05; ***p<0.01

The results happen to be more inline with our hypothesis after adding those subjects. The p-value decreased from 0.28 to 0.032, and our ATE is -13.1, a negative number means the bounding boxes from treatment are more accurate. Controlling using mobile devices as a control, we see much of the variance is explained by the use of mobile devices, though our p-value increases this time when we used this control.

	Dependent variable:			
	bounding_box_score			
	Target Alone	Mobile	Reward	Mobile and Reward
	(1)	(2)	(3)	(4)
in_treatment	−13.050**	−11.100*	−13.014**	−11.089*
	p = 0.032	p = 0.070	p = 0.033	p = 0.071
is_mobile		71.774***		71.536***
		p = 0.000		p = 0.00000
0.20			5.146	1.052
			p = 0.402	p = 0.866
Constant	28.934***	22.947***	26.714***	22.504***
	p = 0.000	p = 0.00000	p = 0.00000	p = 0.00002
Observations	687	660	687	660
R^2	0.007	0.053	0.008	0.053
Adjusted R^2	0.005	0.050	0.005	0.049
Residual Std. Error	79.528 (df = 685)	78.440 (df = 657)	79.545 (df = 684)	78.498 (df = 656)
F Statistic	4.625** (df = 1; 685)	18.399*** (df = 2; 657)	2.663* (df = 2; 684)	12.258*** (df = 3; 656)
Note:				*p<0.1; **p<0.05; ***p<0.01

Covariate Balance Check

	Dependent variable:			
	in_treatment == 0			
	Mobile	Age	Education	Income
	(1)	(2)	(3)	(4)
is_mobile == 1	0.078 (0.078)			
age21to30		0.005 (0.043)		
age31to40		−0.057 (0.069)		
age41to50		−0.119 (0.112)		
agelto21		0.278 (0.169)		
ageover50		0.071 (0.191)		
edufouryearcollege			−0.015 (0.046)	
eduhighschool			−0.206* (0.123)	
edulthighschool			0.000 (0.354)	
edumasterorabove			0.182** (0.079)	
edusomecollege			−0.052 (0.066)	
income10ktolt30k				0.038 (0.058)
incomegt30klt60k				−0.035 (0.060)
incomegt60klt90k				−0.012 (0.080)
incomegt90k				−0.136 (0.110)
incomelt10k				0.028 (0.073)
Constant	0.490*** (0.020)	0.500*** (0.025)	0.500*** (0.025)	0.500*** (0.025)
Observations	670	696	695	695
R^2	0.001	0.007	0.014	0.004
Adjusted R^2	−0.00001	−0.0001	0.007	−0.003
Residual Std. Error	0.500 (df = 668)	0.500 (df = 690)	0.499 (df = 689)	0.501 (df = 689)
F Statistic	0.993 (df = 1; 668)	0.981 (df = 5; 690)	1.924* (df = 5; 689)	0.552 (df = 5; 689)
Note:				*p<0.1; **p<0.05; ***p<0.01

Our covariate balance is well balanced for the most part, the exception being mobile which is off slightly. However this shows that are convenience sampling was an effective method for randomization.

Experiment III - Immediate Payment

In this experiment, we test if the threat of non-payment for poor performance. Our hypothesis was that this would cause similar behavior as the Future Payoff experiment, since the MTurkers would again be able to infer that their performance would be measured only this time there was no additional incentive. The goal was to gain some intuition on how much being tested affected behavior.

Regression Analysis

	Dependent variable:
	bounding_box_score
	Target Alone
in_treatment	−0.685
	p = 0.711
Constant	13.602***
	p = 0.000
Observations	398
R^2	0.0003
Adjusted R^2	−0.002
Residual Std. Error	18.373 (df = 396)
F Statistic	0.138 (df = 1; 396)
Note:	*p<0.1; **p<0.05; ***p<0.01

With 200 subjects in each of treatment and control, we there was a p-value of 0.71, no statistically significant effect with 400 subjects.

Power Test

```
Two-sample t test power calculation

         n = 8876.68
     delta = 0.685011
        sd = 18.35301
 sig.level = 0.05
     power = 0.8
alternative = one.sided
```

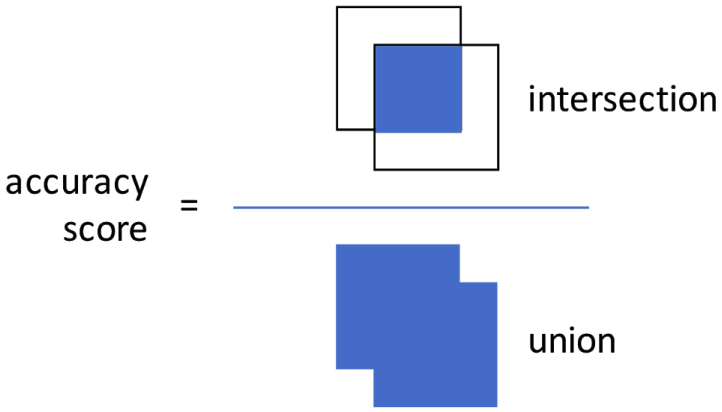
The power test for this treatment suggested we needed 8900 subjects in both treatment and control making finding a statistically significant effect unfeasible, we did not continue down this path.

Experiment IV - Kitchen Sink

After reviewing the results of the first three experiments we hypothesized that one reason we might not be seeing an effect was that the task was too easy. As a result we ran a fourth experiment with N=400 for control and N=400 for treatment, the three motivation types (social, immediate payoff, and future payoff), and a more challenging task of placing 6 bounding boxes, one per car, instead of one. We used the following image for this experiment:

Control average number of bounding boxes drawn per MTurker was 5.69, while for treatment it was 6.04. The tool did not limit the maximum number of bounding boxes. For example, if an MTurker placed a very small bounding box for car 1 by accident and the placed a better box for car 1, then we would have two bounding box entries for car 1 for this worker.

For this final experiment, we added a second accuracy measure called Intersection over Union (IOU) score to see if this would reveal a larger treatment effect size. This score is defined visually [4] in the image below:



When the bounding box at least partially overlaps the ground truth, the IOU score will range from 0 to 1. Due to the complexity of the tool's UI, we decided (post-facto) for this experiment to filter out bad boxes (boxes with a very small area and boxes that were non-overlapping with the ground truth box). In total, we filtered out approximately 200 of the boxes 4750 bounding boxes using this criteria.

Even with this more challenging task, we did not see a statistically significant result. The regression below for the euclidean distance score shows the results are both practically and statistically insignificant with the mean for the control group is 9.27 below the mean for the control group (lower is better) with a p-value of 0.136. Similarly for IOU score we cannot disprove the null hypothesis that the means of the IOU scores are different with an effect of 0.012 and a p-value of 0.14.

Table 27: Subject mean scores, no Controls

	Dependent variable:	
	euclidean_score_filt Euclidean	iou_score_filt Intersection Over Union
	(1)	(2)
Treat	−9.274	0.012
	p = 0.136	p = 0.140
Constant	61.013***	0.888***
	p = 0.000	p = 0.000
Observations	761	761
R^2	0.003	0.003
Adjusted R^2	0.002	0.002
Residual Std. Error (df = 759)	85.646	0.109
F Statistic (df = 1; 759)	2.231	2.183
Note:		*p<0.1; **p<0.05; ***p<0.01

When we control for the covariates (Q1 is question 2 in the Appendix (monitor size) and Q2 is question 3 in the Appendix (mouse-movement device)), we only see a significant p-value of 0.056 for Q1-2 (tablet size monitor) for the Euclidean score. However, considering this p-value in isolation given so many covariates is not proper and one should really adjust it, for example, by using Bonferroni correct.

Table 28: Subject mean scores, with Controls

	Dependent variable:	
	euclidean_score_filt Euclidean	iou_score_filt Intersection Over Union
	(1)	(2)
Treat	−7.259 p = 0.236	0.009 p = 0.267
as.factor(Survey_Q1)1	25.492 p = 0.582	−0.028 p = 0.631
as.factor(Survey_Q1)2	98.530* p = 0.056	−0.098 p = 0.138
as.factor(Survey_Q1)3	−2.526 p = 0.951	0.003 p = 0.953
as.factor(Survey_Q1)4	7.755 p = 0.847	−0.004 p = 0.936
as.factor(Survey_Q1)5	−16.469 p = 0.683	0.028 p = 0.584
as.factor(Survey_Q1)6	102.620 p = 0.102	−0.084 p = 0.293
as.factor(Survey_Q2)1	−2.666 p = 0.935	0.012 p = 0.777
as.factor(Survey_Q2)2	−12.351 p = 0.707	0.025 p = 0.559
as.factor(Survey_Q2)3	58.692 p = 0.118	−0.050 p = 0.299
as.factor(Survey_Q2)4	−41.221 p = 0.646	0.051 p = 0.656
Constant	64.150** p = 0.024	0.870*** p = 0.000
Observations	761	761
R^2	0.063	0.050
Adjusted R^2	0.049	0.036
Residual Std. Error (df = 749)	83.581	0.107
F Statistic (df = 11; 749)	4.574***	3.561***
Note:		*p<0.1; **p<0.05; ***p<0.01

Survey_Q1: How large is your monitor that you use for Mturker HITs? (Survey_Q1)1: Cellphone; (Survey_Q1)2: Tablet Size; (Survey_Q1)3: Small Laptop; (Survey_Q1)4: Midsize Screen; (Survey_Q1)5: Large Screen; (Survey_Q1)6: Not Sure. **Survey_Q2:To move the cursor do you primarily use which of the following?** (Survey_Q2)1: Mouse; (Survey_Q2)2:Trackpad; (Survey_Q2)3: Touch Screen; (Survey_Q2)4:Other.

Conclusion

From our experiments we could not reject our null hypothesis that MTurker workers perform the same with and without a reading a motivational statement. We can see from the experiments that we found neither statistically nor practically significant results. However, there is evidence suggesting that MTurkers are more diligent when they received treatment in Experiment II (Future Payoff). We recommend further exploration here.

Even when we leveraged a more complicated task in Experiment IV, we were not able to see a significant treatment effect. As we did not filter the MTurkers by location and ability, we are unable to make a statement on how our non-results would apply to more specific sub-groups within the MTurker community (e.g. living in the US) nor how workers on other crowdsourcing platforms would respond to our treatments. The one silver lining is for the most part all the MTurker workers did a quality job.

Future Enhancements

There are several avenues to explore to elicit a treatment effect. One is to use a more complicated task such as marking the points on the boundary of an object within an image and/or a task that the MTurkers have likely not seen before. Another is to confirm that the MTurker actually read and understood the motivations. Lastly, a third approach is to come up with new motivations.

We saw that attrition was very high when we recruited via email MTurkers, who earlier completed our survey, to participate in an experiment. This prevented us from being able to capture pre-test data and important covariates via a survey and then do randomized blocking based on this collected pre-data. One way pre-data could be collected is by using a more dynamic tool that collects the pre-data and then right away randomly assigns the Mturker to a control or test group. This would minimize the attrition caused by the time lapse between survey and experiment.

Given we have a few outliers that may significantly affect results, another avenue to explore is using a more robust regression function that can de-weight outlier observations. One such function is rlm (robust regression function) in the MASS package [5].

References

[1] N. Kaufmann et al. *More than fun and money. Worker Motivation in Crowdsourcing--A Study on Mechanical Turk*, Conference: Americas Conference on Information Systems (AMCIS), August 4-7, 2011. Detroit, Michigan

[2] D. Chandler et al. *Breaking Monotony with Meaning: Motivation in Crowdsourcing Markets*, MIT, Oct 3, 2012 (Link)

[3] J. Gancalves et al. *Motivating participation and improving quality of contribution in ubiquitous crowdsourcing*, Computer Networks 90 (2015) 34-48.

[4] *Tutorial: Measuring the accuracy of bounding box image annotations from MTurk*, Amazon Medium Article, Feb 19, 2017 (Link)

[5] R MASS Package, UCLA Institute for Digital Research and Education (Link)

Appendix A - Messaging for Experiments

The pre-treatment survey

1. Have you done any HITs that require putting bounding box on images before? Below is an example: the left side is the original picture (not shown below but in the actual survey), and you are asked to draw a box on the main object of the picture as shown in the right picture.

A. Yes
B. No
C. I don't remember/I don't know.

2. How large is your monitor that you use for Mturker HITs?

A. Cell phone size (6.5" or less)
B. Tablet size (7" to 10")
C. Small laptop size (11" to 13")
D. Mid-size screen (14" to 17")
E. Large screen (18" or larger)
F. I don’t know

3. To move the cursor do you primarily use which of the following?

A. Touch Screen
B. Mouse
C. Trackpad
D. Other

4. What best describes your age group?

A. Under 21
B. 21-30
C. 31-40
D. 41-50
E. Over 50 years old

5. What best describes your annual household income in US dollars?

A. Equal to or below $10,000
B. $10,001 to $30,000
C. $30,001 to $60,000
D. $60,001 to $90,000
E. $90,001 or higher

6. What best describes your education?

A. Less than high school
B. High school graduate
C. Some college
D. 4-year college degree
E. Master of professional degree or higher