Team mTurk - Motivating Quality Work

Kevin Hanna, Kevin Stone, Changjing Zhao

Contents

Motivating Quality Work

What motivates crowdsourced workers to do quality work?

Our scoring metric measures the accuracy of the bounding box by calculating the euclidean distance of the Turkers bounds to the correct bounding. Therefor a **lower score is better**. When the treatment should cause a negative reaction, the score should increase if our hypothesis is correct.

Our Datasets

1. Bound 20 images with negative treatment (Government Surveillance)
2. Bound a single image with negative treatment (Government Surveillance)
3. Bound a single image with positive treatment (Potential future work)
4. Increase subjects for above dataset, 3
5. Bound a single image with negative treatment, reward 2 cents (Threat of not paying for poor performance)

6. Bound a single image with negative treatment, increased reward to 5 cents (Threat of not paying for poor performance)
7. Increase subjects for above datasets 3 & 4 above, smaller reward.

dataset_no	is_pilot	in_treatment	count	mean_score	std_dev
1	1	0	400	136.75944	294.34683
1	1	1	400	135.83448	294.93957
2	1	0	200	15.60004	35.84162
2	1	1	200	19.55064	48.77030
3	0	0	50	18.65368	22.68092
3	0	1	50	25.60954	38.83744
4	0	0	99	39.54211	135.19210
4	0	1	100	14.70005	24.72231
5	0	0	100	13.46661	22.59872
5	0	1	100	11.62734	10.96223
6	0	0	100	13.73527	20.42786
6	0	1	100	14.19297	17.46770
7	0	0	200	26.23898	108.95599
7	0	1	200	14.16740	20.04359

dataset_no	Mean Score	95th pct	Treatment	Control	Total	No Bouding	is_mobile	Reward	Std Dev
1	136.29754	1000.88377	400	400	800	1	0	$0.02	294.45867
2	17.57036	47.17382	200	200	400	3	0	$0.02	42.77251
3	22.02405	87.66010	50	50	100	3	98	$0.20	31.58370
4	27.05803	50.36806	100	99	199	2	191	$0.20	97.49802
5	12.54698	34.39661	100	100	200	2	0	$0.02	17.73936
6	13.96412	55.46328	100	100	200	0	0	$0.05	18.95908
7	20.15711	53.11609	200	200	400	7	381	$0.05	78.18923

experiment	Mean Score	Treatment	Control	Total	Attriters	is_mobile	Std Dev
future	22.39958	350	349	699	12	670	79.73801
immediate	13.25911	200	200	400	2	0	18.35301
pilot	136.29754	400	400	800	1	0	294.45867
social	17.57036	200	200	400	3	0	42.77251

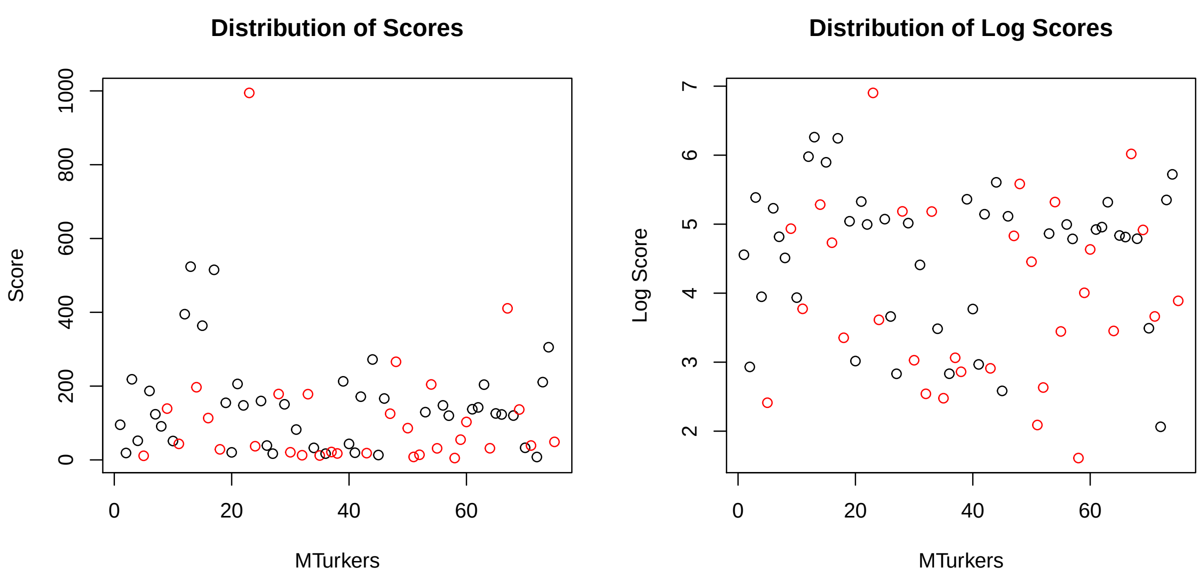
bounding_box_score
Min. : 1.000
1st Qu.: 6.793
Median : 12.645
Mean : 59.861
3rd Qu.: 28.075
Max. :1284.400
NA's :18

1. Pilot

For our pilot, we gave the Turkers a social treatment suggesting their work would be used to improve a government survielance 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.

1.1 EDA



	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

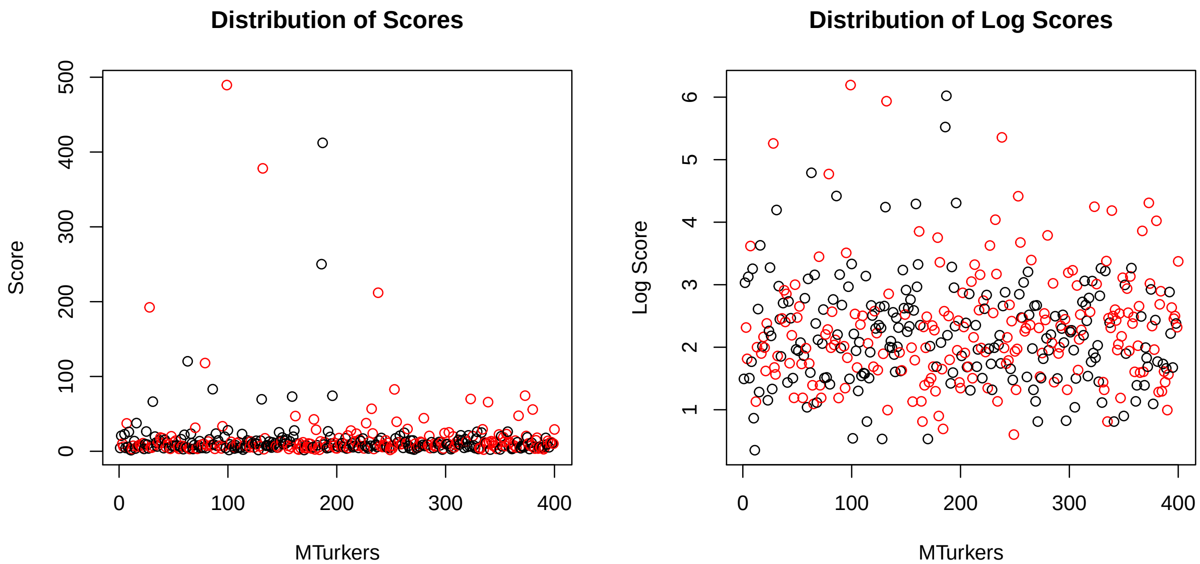
in_treatment	mean_score	std_dev
0	148.0579	124.3664
1	115.7390	187.7335

```
#TODO Gauge if effort decreases with more HITTs
```

2. 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 a single image and created a single bounding box.

2.1 EDA



2.1.1 Score Summary Statistics

Summary Statistics for Score

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

in_treatment	mean_score	std_dev
0	15.60004	35.84162
1	19.55064	48.77030

2.2 Regression Analysis

The results of our regression failed to show any reliable affect of our treatment. The coeffecient is negative, which for our scoring means there is a positive influence from the treatment. But with a p-value of 0.36 there no information can be gleaned from this with any confidence.

With this experiment, the only covariate we had was the amount of time each Turker spent on the task. And working time doesn't seem to be affected by our treatment.

Table 9:

	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

Table 10:

	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.

2.3 Power Test

To achieve the statistical power of 0.8 at the 0.05 confidence-level with the variance we had in this experiement, we would require 1,450 subjects in each 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
##
## NOTE: n is number in *each* group
```

2.4 Learnings from our second 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.

3 Future Payoff

In both of our pilots, we used a treatment which we hypothesized would cause the Turkers in treatment to work less hard, and the ATE was positive, which in our scoring means the bounding was less accurate. We also wanted to test if a positive treatment would have a larger ATE, so the Turkers in treatment were told we were looking for Turkers to perform some future work with the hypothesis that if the Turkers though of the task as a test with the incentive of future work they would try harder. So we ran a small experiment to test this theory.

3.1 Initial Experiment

3.1.1 EDA

3.1.2 Regression Analysis At first look there doesn't seem to be any significant treatment affect, the last 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.

Table 11:

	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

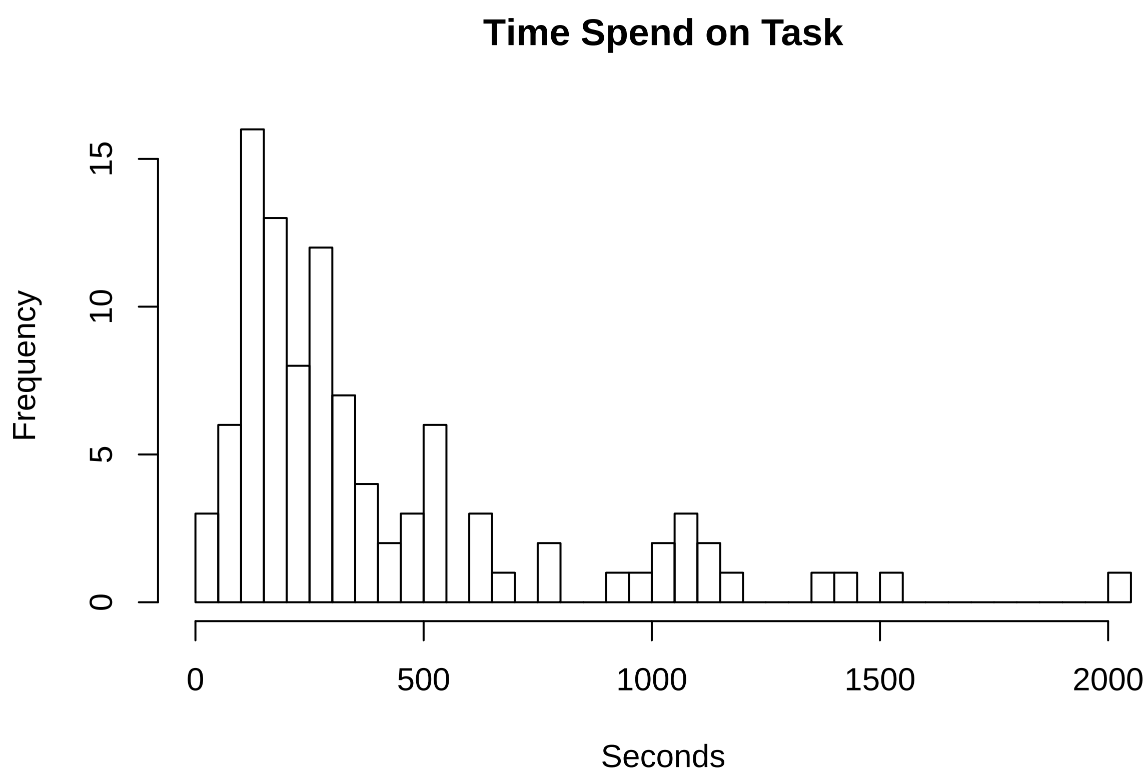
3.1.3 Covariate Regression Analysis In this experiment we asked the Turkers to answer some questions about the device they were using, their experience doing these types of tasks and some demographic info.

The only covariate which seemed to act as any type of control was the education question, though it wasn't very significant. However, all of the coeffecients for the screen size question were negative, and by a fairly significant ammount. The baseline value was cellphone, which is smaller than all the other types of screens. So we tested that on its own.

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

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

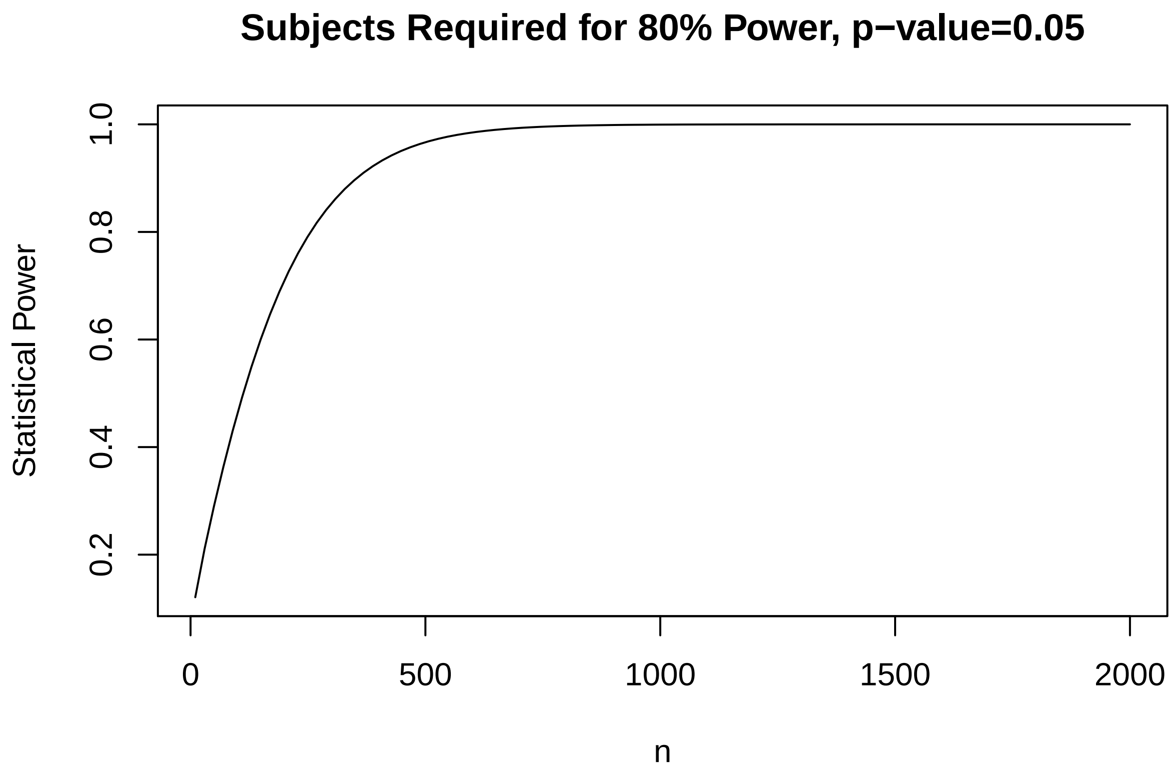
The regression shows that those in our future payoff treatment on average spent 13 seconds more time, the opposite from our previous treatment, which is what we hypothesize, however, the p-values is quite large.



There are alot of values well over the reasonable it should take to perform this task, suggesting that Turkers are not conentrating on our task, it could be they are spawning multiple tabs. Regardless, working time is not helpful for our experiment.

3.1.4 Power Test How much more data would we need?

```
##
##      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
##
## NOTE: n is number in *each* group
```



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 5,800 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.

3.2 Future Payoff - Statistical Power (need better description)

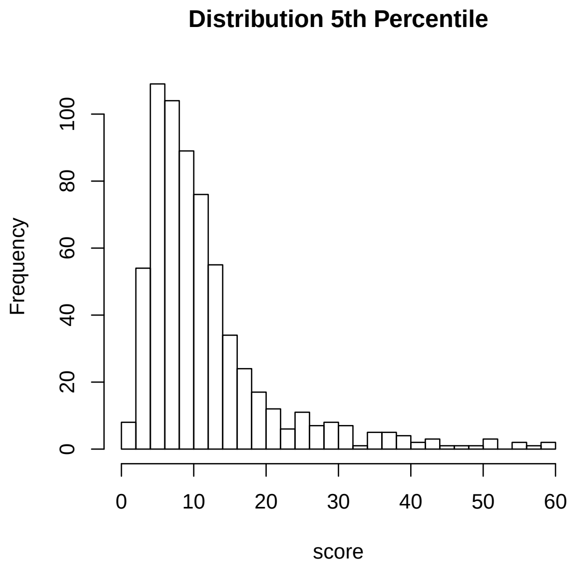
To improve the statistical power for this experiment, we collected data from 600 more subjects.

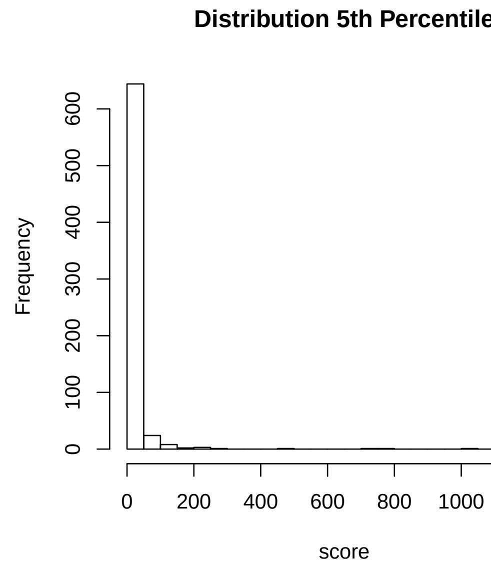
3.2.1

3.2.2 Regression Analysis

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu % Date and time: Sun, Dec 08, 2019 - 11:43:03 PM

The results happend to be more inline with our hypothesis after adding another 600 subjects. The p-value decrease the p-value 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 a much of the variance is explained by the use of mobile devices, though our p-value increases this time when we used this control.





3.2.3 Handling Outliers

3.2.4 Covariate Balance Check

```
##    in_treatment cell not_cell five_cents twenty_cents attriters
## 1:            1   19      319        200          150         6
## 2:            0   25      307        200          149         6
```

3.2.5 Attrition

4 Immediate Payment

4.1 EDA

4.2 Regression Analysis

Call: lm(formula = bounding_box_score ~ in_treatment)

Residuals: Min 1Q Median 3Q Max -12.054 -8.382 -5.425 -0.221 193.261

Coefficients: Estimate Std. Error t value Pr(>|t|)
(Intercept) 13.602 1.302 10.443 <2e-16 *** in_treatment -0.685 1.842 -0.372 0.71
— Signif. codes: 0 '' ***0.001*** '' *0.01* '' 0.05 '.' 0.1 ' ' 1

Residual standard error: 18.37 on 396 degrees of freedom (2 observations deleted due to missingness) Multiple R-squared: 0.0003492, Adjusted R-squared: -0.002175 F-statistic: 0.1383 on 1 and 396 DF, p-value: 0.7102

4.3 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
```

```
##
## NOTE: n is number in *each* group
```

5 Cross Comparison

Table 12: 3.1.3.1

	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

Table 13: 3.1.3.2

	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

Table 14: 3.1.3 2

	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

Table 15: 3.1.3 2

	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

Table 16: 3.2.2 Regression

	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

Table 17: 3.2.2 2

	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

Table 18:

	Dependent variable:		
	bounding_box_score		
	(1)	(2)	(3)
0.05	1.417	1.417	
	p = 0.443	p = 0.798	
0.20		12.828**	
		p = 0.012	
0.20"			12.116***
			p = 0.005
in_treatment	−0.685	−6.458	−6.458
	p = 0.711	p = 0.124	p = 0.124
Constant	12.889***	15.776***	16.488***
	p = 0.000	p = 0.0005	p = 0.00001
Observations	398	692	692
R^2	0.002	0.015	0.015
Adjusted R^2	−0.003	0.011	0.012
Residual Std. Error	18.382 (df = 395)	55.075 (df = 688)	55.037 (df = 689)
F Statistic	0.365 (df = 2; 395)	3.553** (df = 3; 688)	5.303*** (df = 2; 689)
Note:			*p<0.1; **p<0.05; ***p<0.01