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	397	137.60402	295.29711
1	1	1	396	136.39470	296.29412
2	1	0	187	15.25847	36.79851
2	1	1	189	17.09362	42.27343
3	0	0	48	19.02776	23.08104
3	0	1	47	22.71446	36.73351
4	0	0	93	40.35981	139.47131
4	0	1	94	14.51884	25.36227
5	0	0	96	13.55187	23.01864
5	0	1	97	11.61424	11.00214
6	0	0	94	13.56319	20.70507
6	0	1	92	13.15357	17.04807
7	0	0	181	21.52917	98.68055
7	0	1	191	13.17927	16.12633

dataset_no	Mean Score	Treatment	Control	Total	No Bouding	is_mobile	Reward	Std Dev
1	137.00089	396	397	793	1	0	$0.02	295.60836
2	16.17850	189	187	376	3	0	$0.02	39.59534
3	20.79096	47	48	95	3	93	$0.20	30.26851
4	27.36948	94	93	187	2	182	$0.20	100.54784
5	12.57798	97	96	193	2	0	$0.02	17.98909
6	13.36058	92	94	186	0	0	$0.05	18.93443
7	17.20553	191	181	372	7	360	$0.05	69.52288

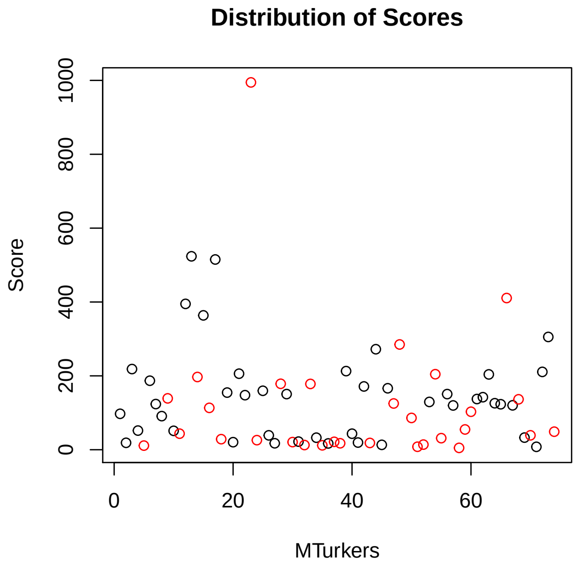
bounding_box_score
Min. : 1.000
1st Qu.: 6.681
Median : 12.414
Mean : 60.752
3rd Qu.: 27.917
Max. :1284.400
NA's :18

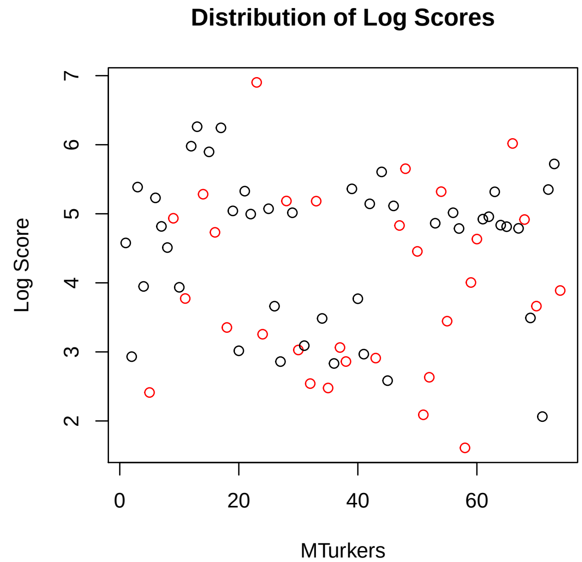
1. Pilot

For our pilot, we gave the Turkers 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 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 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.: 25.938
Median :119.894
Mean :135.288
3rd Qu.:178.160
Max. :994.601
NA's :1

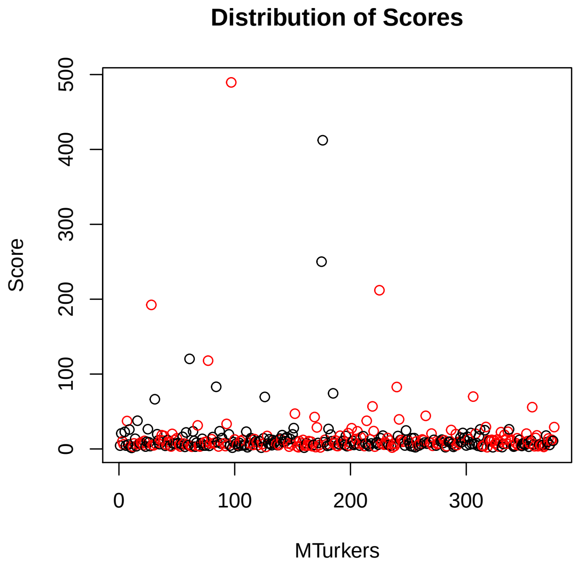
in_treatment	mean_score	std_dev
0	146.7838	125.4300
1	118.8101	190.9985

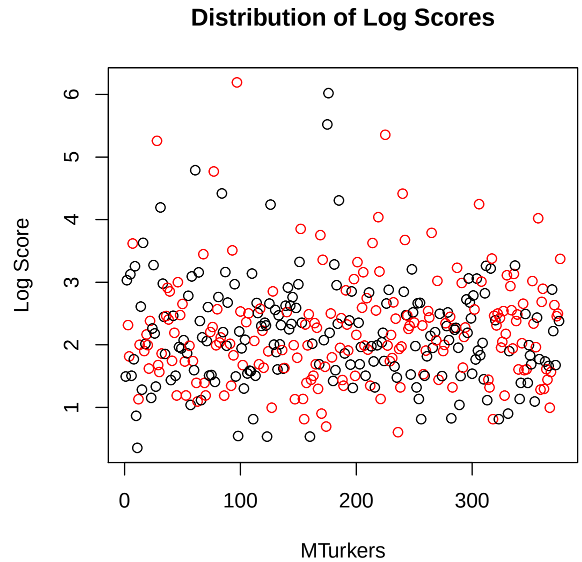
```
#TODO Gauge if effort decreases with more HITTs
```

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.: 5.054
Median : 8.320
Mean : 16.178
3rd Qu.: 13.498
Max. :489.540
NA's :3

in_treatment	mean_score	std_dev
0	15.25847	36.79851
1	17.09362	42.27343

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 0.66 there no information can be gleaned from this with any confidence.

With 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.

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.

Table 8:

	Dependent variable:
	bounding_box_score
in_treatment	1.835 p = 0.656
Constant	15.258*** p = 0.00000
Data Subset	All
Observations	373
R^2	0.001
Adjusted R^2	−0.002
Residual Std. Error	39.638 (df = 371)
F Statistic	0.200 (df = 1; 371)
Note:	*p<0.1; **p<0.05; ***p<0.01

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 nearly 5,800 subjects in both control and treatment.

```
##
##      Two-sample t test power calculation
##
##               n = 5756.986
##           delta = 1.835148
##              sd = 39.59534
##       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 11,600 subjects required to achieve the statistical power we needed was far too many, With a p-value of 0.389, even with the 11,600 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

Table 9:

	Dependent variable:
	WorkTimeInSeconds
in_treatment	−7.720
	p = 0.663
Constant	86.059***
	p = 0.000
Data Subset	All
Observations	376
R^2	0.001
Adjusted R^2	−0.002
Residual Std. Error	171.347 (df = 374)
F Statistic	0.191 (df = 1; 374)
Note:	*p<0.1; **p<0.05; ***p<0.01

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 0.66 in the previous experiment to 0.56 in this, but we only used a quarter the number of subjects.

Table 10:

	Dependent variable:
	bounding_box_score Future Payoff
in_treatment	3.687
	p = 0.563
Constant	19.028***
	p = 0.00004
Data Subset	All
Observations	92
R^2	0.004
Adjusted R^2	−0.007
Residual Std. Error	30.379 (df = 90)
F Statistic	0.338 (df = 1; 90)
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.

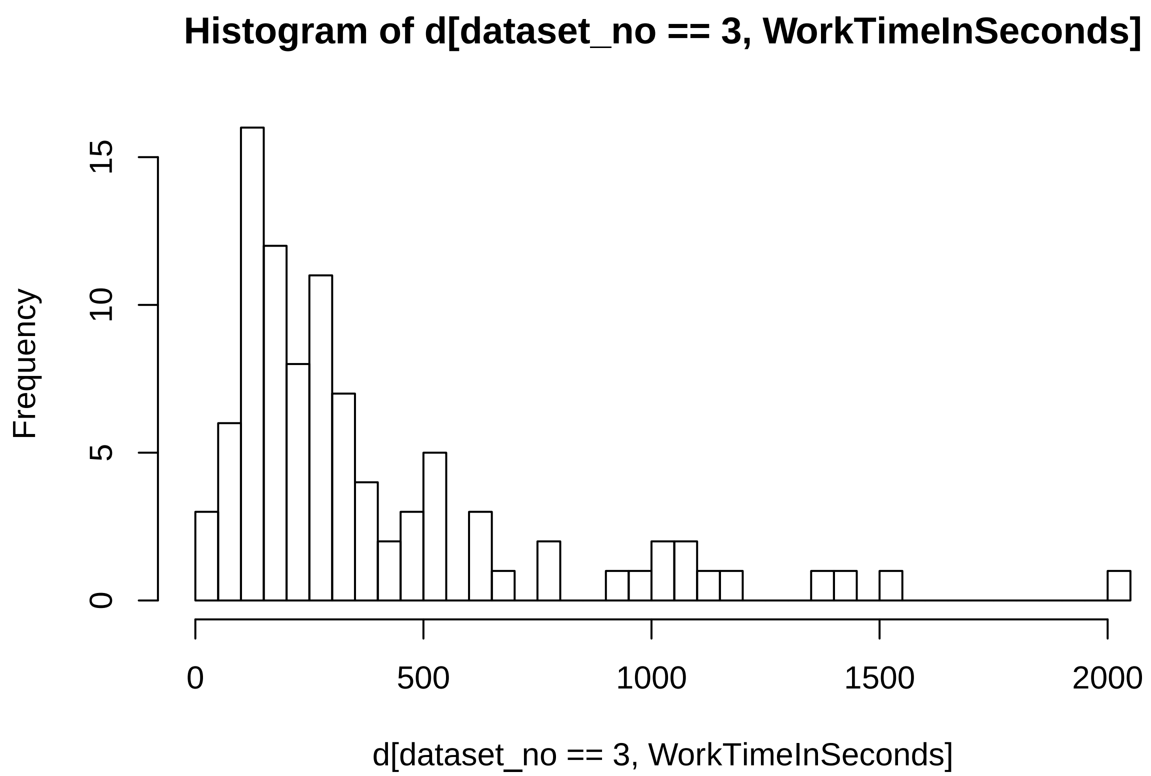
The only covariate which seemed to act as any type of control was the education question, though it wasn't very significant. However, all of the coeffecients for the screensize question were negative, and by a fairly significant ammount. The baseline value was cellphone, which can be significantly smaller 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 0.56 to 0.077. With more data, this could be even lower.

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

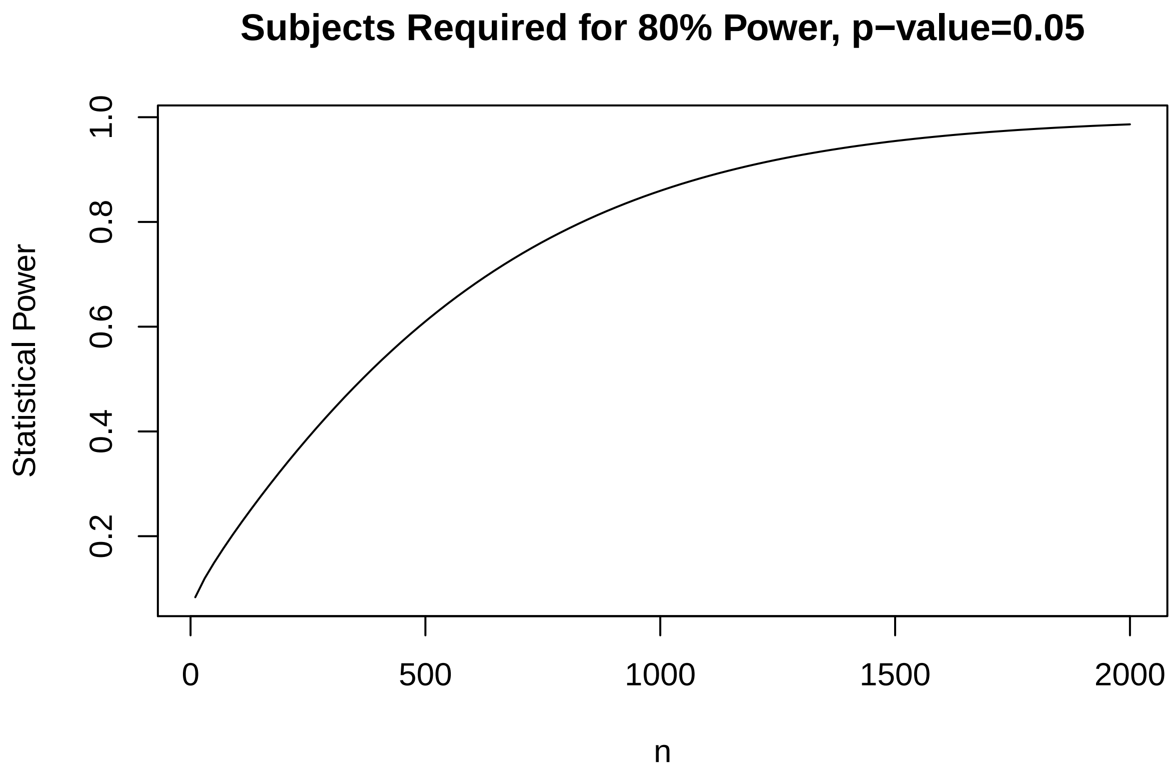
The regression shows those in our future payoff treatment on average spent 23 seconds more time.



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 = 834.1739
##           delta = 3.686703
##              sd = 30.26851
##       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 835 subjects 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: Sat, Dec 07, 2019 - 8:42:35 PM

The results are much better, adding another 700 subjects helped decrease the p-value from 0.56 to 0.05, and our ATE is -11.8, 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 decreased when we used this control.

3.2.3 Handling Outliers

3.2.4 Covariate Balance Check

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 -12.010 -7.997 -5.263 -0.124 193.305

Coefficients: Estimate Std. Error t value Pr(>|t|)
(Intercept) 13.558 1.342 10.099 <2e-16 *** in_treatment -1.190 1.901 -0.626 0.532
— Signif. codes: 0 ‘***’ ***0.001*** ’’ *0.01* ” 0.05 ‘.’ 0.1 ’ ’ 1

Residual standard error: 18.46 on 375 degrees of freedom (2 observations deleted due to missingness) Multiple R-squared: 0.001044, Adjusted R-squared: -0.00162 F-statistic: 0.3918 on 1 and 375 DF, p-value: 0.5317

4.3 Power Test

```
##
##      Two-sample t test power calculation
##
##               n = 2970.269
##           delta = 1.189974
##              sd = 18.4411
##       sig.level = 0.05
##           power = 0.8
##     alternative = one.sided
##
## NOTE: n is number in *each* group
```

5 Cross Comparison

Table 11: 3.1.3 1

	Dependent variable:		
	bounding_box_score		
	Target Alone	Monitor size	Did task before
	(1)	(2)	(3)
in_treatment	7.794	4.708	2.681
	p = 0.231	p = 0.470	p = 0.640
monitorlargescreen	−66.462***		
	p = 0.0004		
monitormidsize	−60.451***		
	p = 0.0005		
monitorsmalllaptop	−57.383***		
	p = 0.002		
monitortablet	−35.061*		
	p = 0.064		
didbfno		7.732	
		p = 0.612	
didbfyes		8.810	
		p = 0.535	
age31to40			−9.611
			p = 0.372
age41to50			97.704***
			p = 0.00001
agelto21			−12.327
			p = 0.653
Constant	72.889***	9.539	18.250***
	p = 0.00004	p = 0.474	p = 0.00003
Observations	92	90	92
R^2	0.200	0.014	0.240
Adjusted R^2	0.153	−0.021	0.205
Residual Std. Error	27.850 (df = 86)	29.643 (df = 86)	26.982 (df = 87)
F Statistic	4.297*** (df = 5; 86)	0.393 (df = 3; 86)	6.879*** (df = 4; 87)
Note:			*p<0.1; **p<0.05; ***p<0.01

Table 12: 3.1.3 2

	Dependent variable:	
	bounding_box_score	
	Education	Income
	(1)	(2)
in_treatment	$-$ 3.280 p = 0.613	-0.015 p = 0.999
eduhighschool	-17.099 p = 0.438	
edumasterorabove	-14.161 p = 0.154	
edusomecollege	-15.051 p = 0.216	
incomegt30klt60k		7.338 p = 0.371
incomegt60klt90k		-3.268 p = 0.762
incomegt90k		19.021 p = 0.160
incomelt10k		-8.956 p = 0.404
Constant	22.433^{***} p = 0.00001	18.375^{***} p = 0.002
Observations	92	92
R^2	0.045	0.056
Adjusted R^2	0.001	0.001
Residual Std. Error	30.255 (df = 87)	30.251 (df = 86)
F Statistic	1.021 (df = 4; 87)	1.021 (df = 5; 86)
Note:		*p<0.1; **p<0.05; ***p<0.01

Table 13: 3.1.3 2

	Dependent variable:	
	bounding_box_score	
	Target Alone	Mobile Control
	(1)	(2)
in_treatment	3.687	11.196*
	p = 0.563	p = 0.077
is_mobile		34.558***
		p = 0.0002
Constant	19.028***	10.296**
	p = 0.00004	p = 0.031
Data Subset	All	All
Observations	92	90
R^2	0.004	0.160
Adjusted R^2	−0.007	0.141
Residual Std. Error	30.379 (df = 90)	28.327 (df = 87)
F Statistic	0.338 (df = 1; 90)	8.291*** (df = 2; 87)
Note:		*p<0.1; **p<0.05; ***p<0.01

Table 14: 3.1.3 2

	Dependent variable:	
	WorkTimeInSeconds	
	Future Payoff	Social
	(1)	(2)
in_treatment	22.983	−7.720
	p = 0.766	p = 0.663
Constant	377.208***	86.059***
	p = 0.000	p = 0.000
Data Subset	All	All
Observations	95	376
R^2	0.001	0.001
Adjusted R^2	−0.010	−0.002
Residual Std. Error	374.924 (df = 93)	171.347 (df = 374)
F Statistic	0.089 (df = 1; 93)	0.191 (df = 1; 374)
Note:		*p<0.1; **p<0.05; ***p<0.01

Table 15: 3.2.2 Regression

	Dependent variable:	
	bounding_box_score	
	n=700	n=100
	(1)	(2)
in_treatment	-11.783^{**}	3.687
	p = 0.050	p = 0.563
Constant	26.632^{***}	19.028^{***}
	p = 0.000	p = 0.00004
Data Subset	All	All
Observations	642	92
R^2	0.006	0.004
Adjusted R^2	0.004	-0.007
Residual Std. Error	75.966 (df = 640)	30.379 (df = 90)
F Statistic	3.861^{**} (df = 1; 640)	0.338 (df = 1; 90)
Note:		*p<0.1; **p<0.05; ***p<0.01

Table 16: 3.2.2 2

	Dependent variable:			
	bounding_box_score			
	Target Alone	Controlling for Mobile	Reward	Mobile and Reward
	(1)	(2)	(3)	(4)
in_treatment	-11.783^{**}	-11.238^{*}	-11.607^{*}	-11.198^{*}
	p = 0.050	p = 0.058	p = 0.054	p = 0.059
is_mobile		81.852^{***}		81.420^{***}
		p = 0.000		p = 0.000
0.20			7.709	1.458
			p = 0.204	p = 0.810
Constant	26.632^{***}	21.251^{***}	23.216^{***}	20.627^{***}
	p = 0.000	p = 0.00000	p = 0.00001	p = 0.00005
Data Subset	All	All	$x == 1$	
Observations	642	625	642	625
R^2	0.006	0.071	0.009	0.071
Adjusted R^2	0.004	0.068	0.005	0.067
Residual Std. Error	75.966 (df = 640)	73.872 (df = 622)	75.929 (df = 639)	73.928 (df = 621)
F Statistic	3.861^{**} (df = 1; 640)	23.772^{***} (df = 2; 622)	2.744^{*} (df = 2; 639)	15.844^{***} (df = 3; 621)
Note:				*p<0.1; **p<0.05; ***p<0.01

Table 17:

	Dependent variable:		
	bounding_box_score		
	Reward		
	(1)	(2)	(3)
0.05	0.773 p = 0.685	0.723 p = 0.901	
0.20		12.547** p = 0.019	
0.20"			12.190*** p = 0.007
in_treatment	−1.184 p = 0.535	−7.414* p = 0.094	−7.418* p = 0.093
Constant	13.173*** p = 0.000	16.305*** p = 0.0005	16.663*** p = 0.00001
Data Subset	All	All	$x == 1$
Observations	377	654	654
R^2	0.001	0.016	0.016
Adjusted R^2	−0.004	0.011	0.013
Residual Std. Error	18.477 (df = 374)	56.365 (df = 650)	56.323 (df = 651)
F Statistic	0.278 (df = 2; 374)	3.451** (df = 3; 650)	5.177*** (df = 2; 651)
Note:			*p<0.1; **p<0.05; ***p<0.01