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 Experiments

1. Pilot - Bound 20 images with negative treatment (Government Surveillance)
2. Pilot - Bound a single image with negative treatment (Government Surveillance)
3. Experiment - Bound a single image with positive treatment (Potential future work)
4. Experiment - Increase subjects for experiment 3
5. Experiment - Bound a single image with negative treatment, reward 2 cents (Threat of not paying for poor performance)
6. Experiment - Bound a single image with negative treatment, increased reward to 5 cents (Threat of not paying for poor performance)

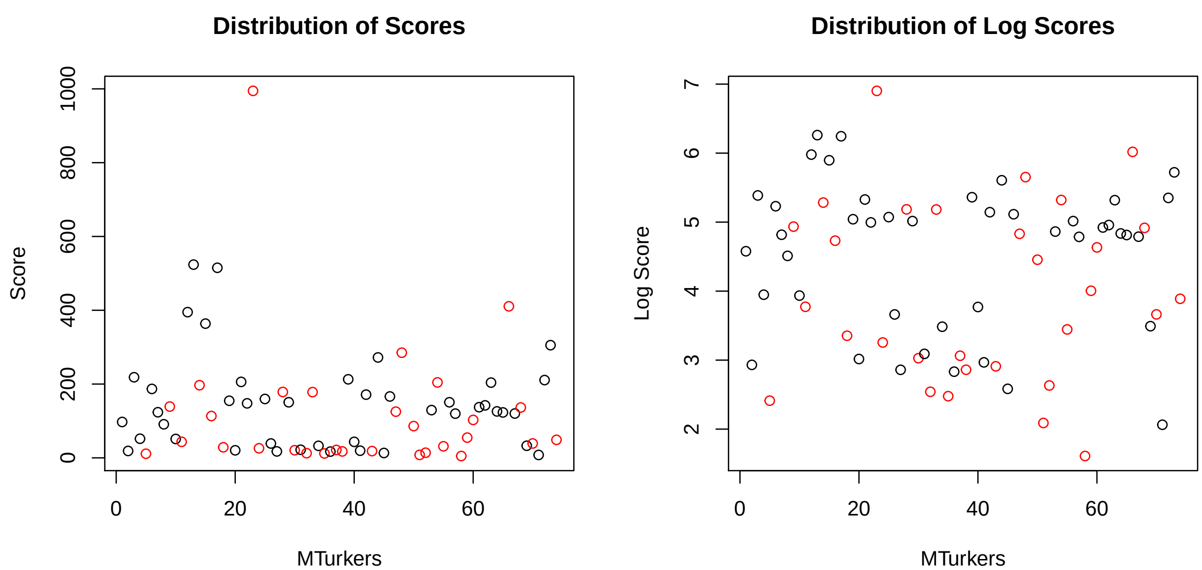
experiment_no	is_pilot	in_treatment	count	mean_score	std_dev
1	1	0	397	137.60402	295.29711
1	1	1	396	136.39470	296.29412
2	1	0	187	15.25847	36.79851
2	1	1	189	17.09362	42.27343
3	0	0	48	19.02776	23.08104
3	0	1	47	22.71446	36.73351
4	0	0	93	40.35981	139.47131
4	0	1	94	14.51884	25.36227
5	0	0	96	13.55187	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	191	13.17927	16.12633
7	0	1	181	21.52917	98.68055

	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

Experiment 1, our first 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.



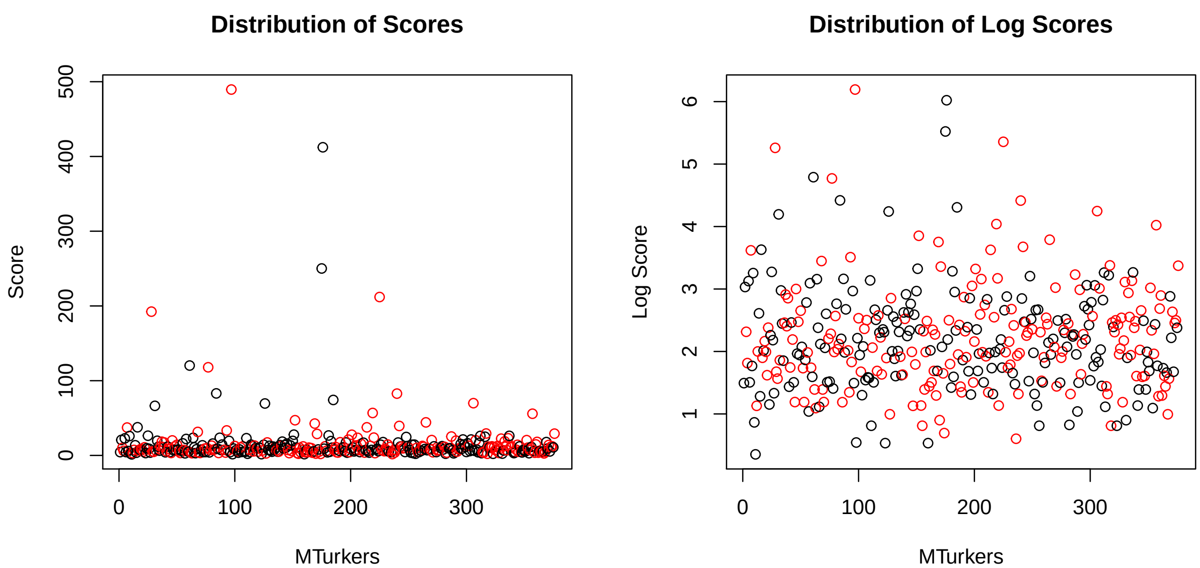
	mean_worker_score
	Min. : 5.003
	1st Qu.: 25.938
	Median :119.894
	Mean :135.288
	3rd Qu.:178.160
	Max. :994.601
	NA's :1

in_treatment	mean_score	std_dev
0	146.7838	125.4300
1	118.8101	190.9985

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

Experiment 2, our second pilot

With the first pilot behind us, we decided we needed to focus on increasing our statistical power and hypothesized that having more subjects with fewer experiments would provide more statistical power.



2.1 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

Our coefficient for in treatment was still more likely due to random noise than not.

2.1 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
```

2.2 Analysis With this pilot, the only covariate we had was the amount of time each Turker spent on the task. And working time acts as a control explaining away some of the variance reducing the p-value for our target feature from 0.45 to 0.39.

Table 1:

	Dependent variable:	
	bounding_box_score	
	Target Alone	With WorkInSeconds Control
	(1)	(2)
in_treatment	1.835	1.903
	p = 0.656	p = 0.644
WorkTimeInSeconds		0.010
		p = 0.426
Constant	15.258***	14.441***
	p = 0.00000	p = 0.00001
Data Subset	All	All
Observations	373	373
R^2	0.001	0.002
Adjusted R^2	−0.002	−0.003
Residual Std. Error	39.638 (df = 371)	39.657 (df = 370)
F Statistic	0.200 (df = 1; 371)	0.418 (df = 2; 370)
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 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.

Experiment 3, incentive of future work

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 Simple regression analysis The first look was disappointing, the last p-value had gone up from 0.45 in the previous experiment to 0.563 in this, but we only used a quarter the number of subjects. More concerning is the change in the treatment was estimated to change the direction of the coeffecient, and it is still positive.

Table 2:

	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

3.2 Analysis with covariates 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.

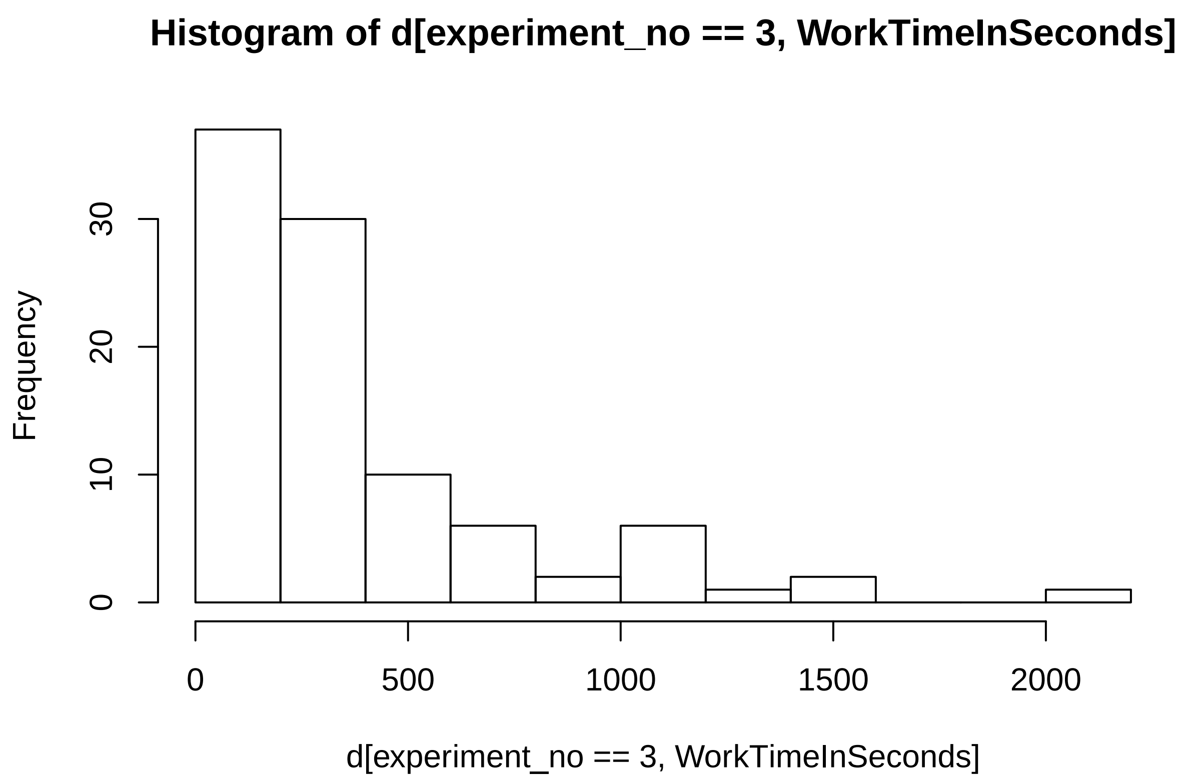
If the subject is using a cellphone to do the task, their accuracy goes down (score increases), which is intuitive. Having cellphone as a control decreases the p-value from 0.56 to 0.33. With more data, this could be quite a bit lower. But it still doesn't explain why subjects with more incentive are doing a poorer job.

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 treatment on average spent 23 seconds more time, this alone is concerning, as the task itself shouldn't take that much time.

Table 3:

	Dependent variable:	
	bounding_box_score	
	Incentivized	Negative Treatment
	(1)	(2)
in_treatment	3.687	1.835
	p = 0.563	p = 0.656
Constant	19.028***	15.258***
	p = 0.00004	p = 0.00000
Data Subset	All	All
Observations	92	373
R^2	0.004	0.001
Adjusted R^2	−0.007	−0.002
Residual Std. Error	30.379 (df = 90)	39.638 (df = 371)
F Statistic	0.338 (df = 1; 90)	0.200 (df = 1; 371)
Note:		*p<0.1; **p<0.05; ***p<0.01



There are alot of values 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 Power Test With a lot of speculation about whether our statistical significance would go up with more data, we tested that theory by doing a power calculation.

```
##
##      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.

Experiment 4, More data

To improve the statistical power from Experiment 3, we are adding more data and sending out another 100 control tasks to Turkers and 100 with the same treatment.

TODO covariate balance check, demographic info show how random it is.

Foo

The results are much better, adding another 200 subjects helped decrease the p-value from 0.56 to 0.12, and our ATE is -15.9, a negative number means the bounding boxes from treatment are more accurate.

Call: lm(formula = bounding_box_score ~ in_treatment + is_mobile + tried)

Residuals: Min 1Q Median 3Q Max -74.74 -17.33 -8.53 -2.23 930.51

Coefficients: Estimate Std. Error t value Pr(>|t|)
(Intercept) 25.475 7.167 3.554 0.000447 * **in_treatment -12.460 9.825 -1.268 0.205791**
is_mobile 51.518 16.535 3.116 0.002031 tried 235.685 81.783 2.882 0.004268 ** — Signif. codes: 0 '' ***0.001*** '' *0.01* '' 0.05 '.' 0.1 ' ' 1

Residual standard error: 81.47 on 273 degrees of freedom (5 observations deleted due to missingness) Multiple R-squared: 0.06902, Adjusted R-squared: 0.05879 F-statistic: 6.747 on 3 and 273 DF, p-value: 0.0002094

```
#e4_mod_3 <- d[experiment_no %in% c(3,4), lm(bounding_box_score ~ in_treatment+factor(monitor))]
#e4_mod_4 <- d[experiment_no %in% c(3,4), lm(bounding_box_score ~ in_treatment+factor(didbf))]
#e4_mod_5 <- d[experiment_no %in% c(3,4), lm(bounding_box_score ~ in_treatment+factor(age))]
#e4_mod_6 <- d[experiment_no %in% c(3,4), lm(bounding_box_score ~ in_treatment+factor(edu))]
#e4_mod_7 <- d[experiment_no %in% c(3,4), lm(bounding_box_score ~ in_treatment+factor(income))]

#stargazer(e4_mod_3, e4_mod_4, e4_mod_5, e4_mod_6, e4_mod_7,
#          type = 'text', header = FALSE, table.placement = 'h',  report=('vc*p'),
#          add.lines = list(c("Data Subset", "All", "All", "$x==1$")))
```

```
e4_ate = d[experiment_no %in% c(3, 4) & in_treatment == 1, mean(bounding_box_score, na.rm=T)] - d[exper

e4_sd = d[experiment_no %in% c(3, 4), sd(bounding_box_score, na.rm=T)]
```

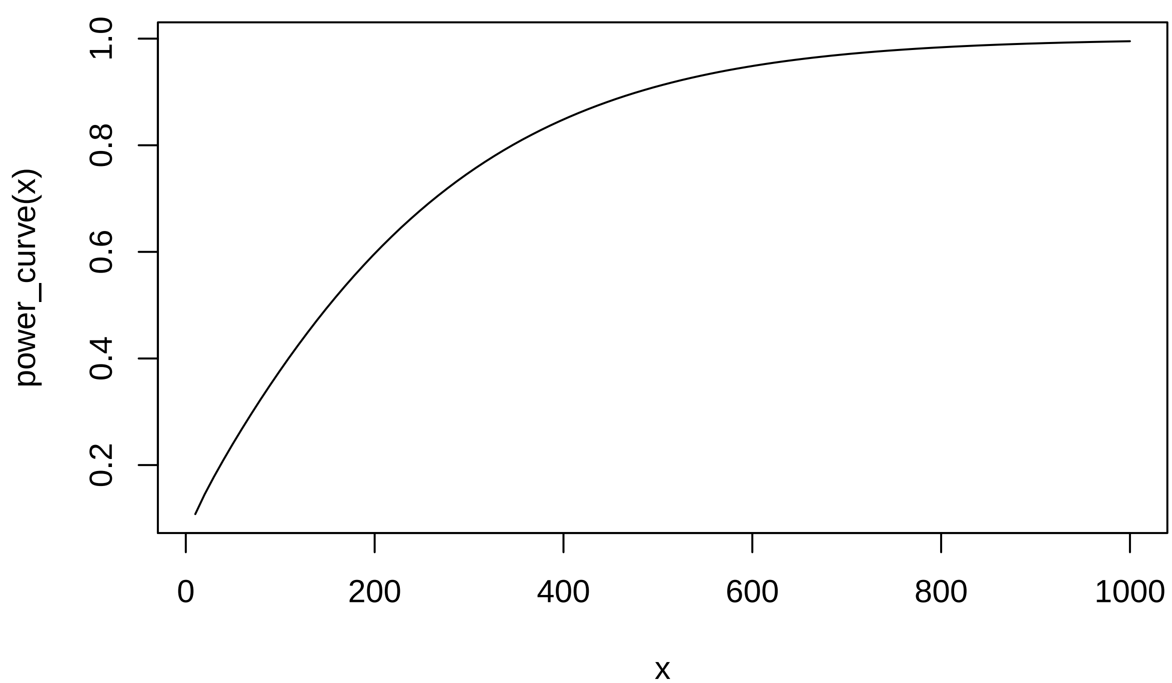
```r
power.t.test(delta=abs(e4_ate),
             sd=e4_sd,
             sig.level = 0.05,
             power = 0.80,
             alternative = "one.sided",
             n = NULL)
```

4.1 Power Test

```
##
##      Two-sample t test power calculation
##
##               n = 345.7973
##           delta = 15.89495
##              sd = 83.97393
##       sig.level = 0.05
##           power = 0.8
##     alternative = one.sided
##
## NOTE: n is number in *each* group
```

```r
power_curve <- function(x) {
  result = c()

  for (i in 1:length(x)) {
    new_n <- power.t.test(delta=abs(e4_ate),
             sd=e4_sd,
             sig.level = 0.05,
             power = NULL,
             alternative = "one.sided",
             n = x[i])["power"]

    result <- c(result, new_n)
  }

  return(result)
}

sig_curve <- function(x) {
  result = c()

  for (i in 1:length(x)) {
    new_n <- power.t.test(delta=abs(e4_ate),
             sd=e4_sd,
             sig.level = NULL,
             power = 0.8,
             alternative = "one.sided",
             n = x[i])["sig.level"]

    result <- c(result, new_n)
  }

  return(result)
}
```

```r
delta_curve <- function(x) {
  result = c()

  for (i in 1:length(x)) {
    new_n <- power.t.test(delta=x[i],
             sd=e4_sd,
             sig.level = 0.05,
             power = 0.8,
             alternative = "one.sided",
             n = NULL)["n"]

    result <- c(result, new_n)
  }

  return(result)
}
curve(power_curve(x), 10, 1000)
```




```r
#curve(sig_curve(x), 10, 1000)
#curve(delta_curve(x), 5, 20)
```

Experiment 5, threats don't work

```r
e5_mod_1 <- d[experiment_no == 5, lm(bounding_box_score ~ in_treatment)]
summary(e5_mod_1)
```

```
##
## Call:
## lm(formula = bounding_box_score ~ in_treatment)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -12.005  -7.388  -4.123   0.854 193.311
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    13.552      1.848   7.334 6.38e-12 ***
## in_treatment   -1.938      2.606  -0.743    0.458
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.01 on 189 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:  0.002916,   Adjusted R-squared:  -0.00236
## F-statistic: 0.5527 on 1 and 189 DF,  p-value: 0.4582
```

Experiment 6, threats still don't work

```
e6_mod_1 <- d[experiment_no == 6, lm(bounding_box_score ~ in_treatment)]
summary(e6_mod_1)
```

```
##
## Call:
## lm(formula = bounding_box_score ~ in_treatment)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -12.02  -8.64  -6.39  -1.38 107.83
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   13.5632     1.9581   6.927 6.99e-11 ***
## in_treatment  -0.4096     2.7842  -0.147    0.883
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.98 on 184 degrees of freedom
## Multiple R-squared:  0.0001176,  Adjusted R-squared:  -0.005317
## F-statistic: 0.02164 on 1 and 184 DF,  p-value: 0.8832
```

```
e6_mod_2 <- d[experiment_no %in% c(5,6), lm(bounding_box_score ~ in_treatment+(Reward == "$0.05"))]
summary(e6_mod_2)
```

```
##
## Call:
## lm(formula = bounding_box_score ~ in_treatment + (Reward == "$0.05"))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -12.399  -8.042  -5.148  -0.011 193.690
##
```

```
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            13.1730     1.6439   8.013 1.44e-14 ***
## in_treatment           -1.1838     1.9033  -0.622    0.534
## Reward == "$0.05"TRUE   0.7731     1.9034   0.406    0.685
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.48 on 374 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:  0.001484,   Adjusted R-squared:  -0.003855
## F-statistic: 0.278 on 2 and 374 DF,  p-value: 0.7575
```

Experiment 7, Even MORE incentive data

```
##
##      Two-sample t test power calculation
##
##               n = 17560.84
##           delta = 2.020332
##              sd = 76.13563
##       sig.level = 0.05
##           power = 0.8
##     alternative = one.sided
##
## NOTE: n is number in *each* group
```

Table 4:

	Dependent variable:			
	bounding_box_score			
	Target Alone	Monitor size	Did task before	Age
	(1)	(2)	(3)	(4)
in_treatment	3.687 p = 0.563	7.794 p = 0.231	4.708 p = 0.470	2.681 p = 0.640
as.factor(monitor)largescreen		-66.462^{***} p = 0.0004		
as.factor(monitor)midsize		-60.451^{***} p = 0.0005		
as.factor(monitor)smalllaptop		-57.383^{***} p = 0.002		
as.factor(monitor)tablet		-35.061^{*} p = 0.064		
as.factor(didbf)no			7.732 p = 0.612	
as.factor(didbf)yes			8.810 p = 0.535	
as.factor(age)31to40				-9.611 p = 0.372
as.factor(age)41to50				97.704^{***} p = 0.00001
as.factor(age)lt21				-12.327 p = 0.653
as.factor(edu)highschool				
as.factor(edu)masterorabove				
as.factor(edu)somecollege				
as.factor(income)gt30klt60k				
as.factor(income)gt60klt90k				
as.factor(income)gt90k				
as.factor(income)lt10k				
Constant	19.028^{***} p = 0.00004	72.889^{***} p = 0.00004	9.539 p = 0.474	18.250^{***} p = 0.00003

Table 5:

	Dependent variable:	
	bounding_box_score	
	Target Alone	Used Cellphone
	(1)	(2)
in_treatment	3.687	2.239
	p = 0.563	p = 0.710
monitor == "cellphone"		58.789***
		p = 0.001
Constant	19.028***	17.803***
	p = 0.00004	p = 0.00005
Data Subset	All	All
Observations	92	92
R^2	0.004	0.123
Adjusted R^2	−0.007	0.104
Residual Std. Error	30.379 (df = 90)	28.655 (df = 89)
F Statistic	0.338 (df = 1; 90)	6.268*** (df = 2; 89)
Note:		*p<0.1; **p<0.05; ***p<0.01

Table 6:

	Dependent variable:	
	WorkTimeInSeconds	
	Incentivized	Negative Treatment
	(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 7:

	Dependent variable:	
	bounding_box_score	
	n=100 + n=200	n=100
	(1)	(2)
in_treatment	−15.895	3.687
	p = 0.116	p = 0.563
Constant	33.046***	19.028***
	p = 0.00001	p = 0.00004
Data Subset	All	All
Observations	277	92
R^2	0.009	0.004
Adjusted R^2	0.005	−0.007
Residual Std. Error	83.748 (df = 275)	30.379 (df = 90)
F Statistic	2.494 (df = 1; 275)	0.338 (df = 1; 90)
Note:		*p<0.1; **p<0.05; ***p<0.01

Table 8:

	Dependent variable:	
	bounding_box_score	
	Target Alone	Cellphone
	(1)	(2)
in_treatment	−15.895	−14.181
	p = 0.116	p = 0.155
is_mobile		50.409***
		p = 0.003
Constant	33.046***	27.285***
	p = 0.00001	p = 0.0002
Data Subset	All	All
Observations	277	277
R^2	0.009	0.041
Adjusted R^2	0.005	0.034
Residual Std. Error	83.748 (df = 275)	82.547 (df = 274)
F Statistic	2.494 (df = 1; 275)	5.812*** (df = 2; 274)
Note:		*p<0.1; **p<0.05; ***p<0.01

Table 9:

	Dependent variable:			
	bounding_box_score			
	n=300	n=300 and cellphone	n=700	n=700 and cellphone
	(1)	(2)	(3)	(4)
in_treatment	−15.895	−14.181	−2.020	−16.466
	p = 0.116	p = 0.155	p = 0.738	p = 0.105
is_mobile		50.409^{***}		
		p = 0.003		
is_cellphone				25.693
				p = 0.426
Constant	33.046^{***}	27.285^{***}	21.633^{***}	32.679^{***}
	p = 0.00001	p = 0.0002	p = 0.00000	p = 0.00001
Data Subset	All	All	$x == 1$	
Observations	277	277	642	277
R^2	0.009	0.041	0.0002	0.011
Adjusted R^2	0.005	0.034	−0.001	0.004
Residual Std. Error	83.748 (df = 275)	82.547 (df = 274)	76.188 (df = 640)	83.803 (df = 274)
F Statistic	2.494 (df = 1; 275)	5.812^{***} (df = 2; 274)	0.113 (df = 1; 640)	1.565 (df = 2; 274)
Note:				*p<0.1; **p<0.05; ***p<0.01