You Get What You Pay For

Experimental Analysis on the Relationship Between Reward and Productivity

W241-1 Experiments and Causality
Legg Yeung, Stanimir Vichev, Frederic Suares
12/13/2018

Hypothesis

Higher Reward (Treatment)

Pay rate per HIT [$0.10 - $0.55]

Better Productivity (Outcome)

Returned accuracy per HIT [0%-100%]

Our Workflow

Sort by: HITs Available (most first) GO! Show all details | Hide all details 1 2 3 4 5 Next Items per Page: 10

Extract purchased items from a shopping receipt (1-2 items) — View a HIT in this group
Requester: ScoutIt | HIT Expiration Date: Dec 20, 2017 (6 days 23 hours) | Reward: $0.01
Time Allotted: 2 hours | HITs Available: 120668

Extract purchased items from a shopping receipt (3-5 items) — View a HIT in this group
Requester: ScoutIt | HIT Expiration Date: Dec 20, 2017 (6 days 23 hours) | Reward: $0.03
Time Allotted: 2 hours | HITs Available: 102909

Extract purchased items from a shopping receipt (3-5 items) — View a HIT in this group
Requester: ScoutIt | HIT Expiration Date: Dec 20, 2017 (6 days 20 hours) | Reward: $0.04
Time Allotted: 2 hours | HITs Available: 27542

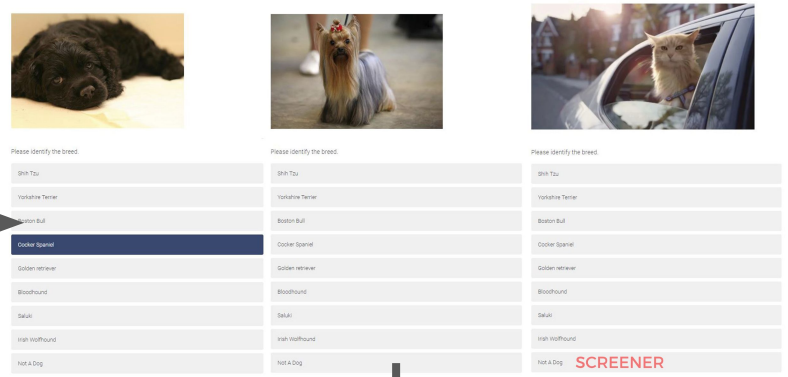
Extract purchased items from a shopping receipt (6-10 items) — View a HIT in this group
Requester: ScoutIt | HIT Expiration Date: Dec 20, 2017 (6 days 21 hours) | Reward: $0.07
Time Allotted: 2 hours | HITs Available: 22324

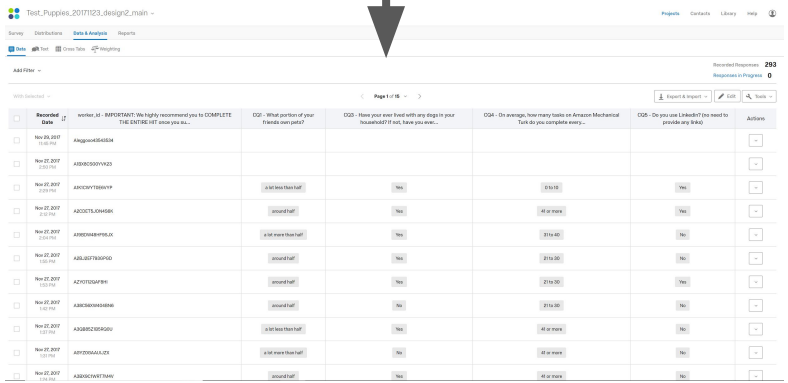
Facial Attributes Annotation — Request Qualification (Why?) | View a HIT in this group
Requester: Michele Merler | HIT Expiration Date: Jan 30, 2018 (6 weeks 6 days) | Reward: $0.10
Time Allotted: 60 minutes | HITs Available: 22215

Locate Box, Label, Barcode, and Tracking Code in Image — View a HIT in this group
Requester: Tony Nguyen | HIT Expiration Date: Dec 19, 2017 (6 days 9 hours) | Reward: $0.02
Time Allotted: 5 minutes | HITs Available: 19954

Extract purchased items from a shopping receipt — View a HIT in this group
Requester: ScoutIt | HIT Expiration Date: Dec 20, 2017 (6 days 22 hours) | Reward: $0.09
Time Allotted: 2 hours | HITs Available: 19790

Extract purchased items from a shopping receipt (1-2 items) — View a HIT in this group
Requester: ScoutIt | HIT Expiration Date: Dec 20, 2017 (6 days 20 hours) | Reward: $0.02
Time Allotted: 2 hours | HITs Available: 13990





Covariate Questions

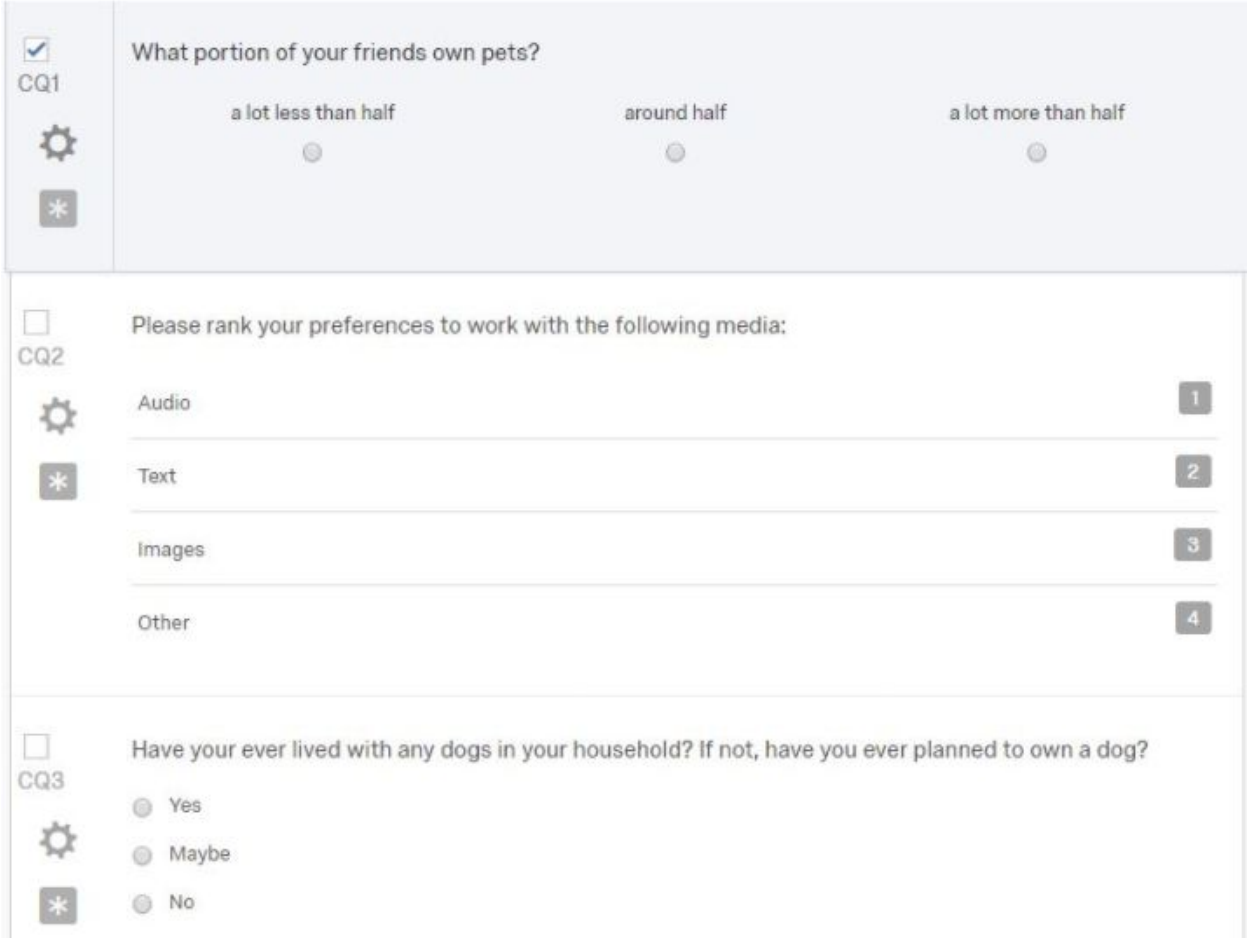


Image Classification Questions

Please identify the breed.

- Shih Tzu
- Yorkshire Terrier
- Boston Bull
- **Cocker Spaniel**
- Golden retriever
- Bloodhound
- Saluki
- Irish Wolfhound
- Not A Dog

Please identify the breed.

- Shih Tzu
- Yorkshire Terrier
- Boston Bull
- Cocker Spaniel
- Golden retriever
- Bloodhound
- Saluki
- Irish Wolfhound
- Not A Dog

Please identify the breed.

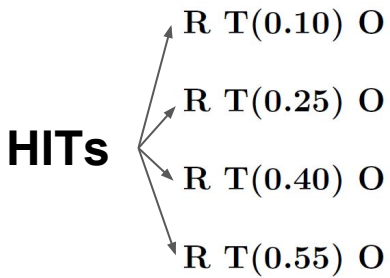
- Shih Tzu
- Yorkshire Terrier
- Boston Bull
- Cocker Spaniel
- Golden retriever
- Bloodhound
- Saluki
- Irish Wolfhound
- Not A Dog

SCREENER

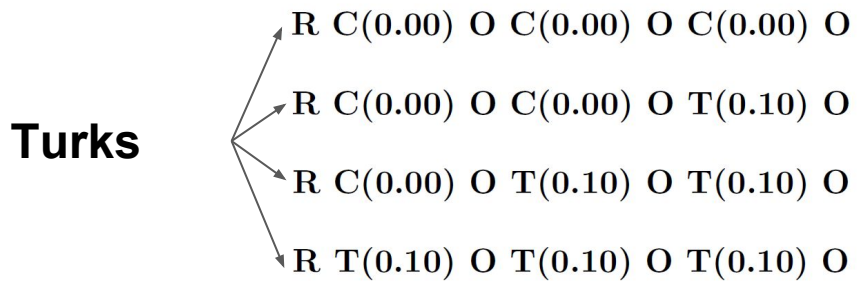
(Images sourced from Standford Dog Dataset)

Experiment Designs

Design 1 Notation: Between Subject Design



Design 2 Notation: Stepped-Wedge Design



Experiment Schedules

Experiment Schedule for Design 1

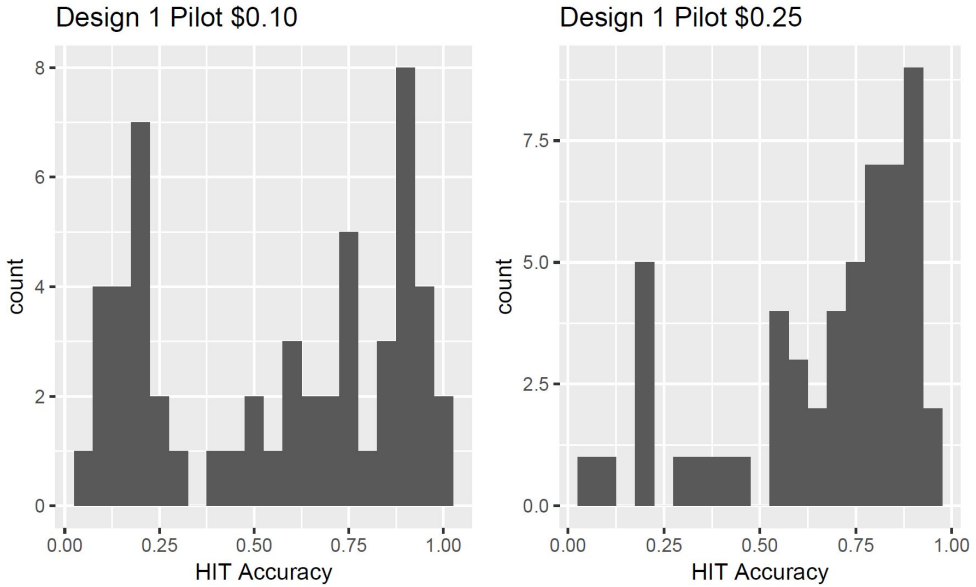
Publish Order	Date	Time Frame	Treatment (Pay Rate)	Available HITs
Pilot	Oct 28, 2017 (Saturday)	Morning	$0.10	50
Pilot	Oct 29, 2017 (Sunday)	Afternoon	$0.25	50
1	Nov 11, 2017 (Saturday)	Morning	$0.10	100
1	Nov 11, 2017 (Saturday)	Afternoon	$0.55	100
1	Nov 12, 2017 (Sunday)	Morning	$0.25	100
1	Nov 12, 2017 (Sunday)	Afternoon	$0.40	100
2	Nov 18, 2017 (Saturday)	Morning	$0.40	100
2	Nov 18, 2017 (Saturday)	Afternoon	$0.25	100
2	Nov 19, 2017 (Sunday)	Morning	$0.55	100
2	Nov 19, 2017 (Sunday)	Afternoon	$0.10	100

Experiment Schedule for Design 2

Name	Date	Time Frame	Base pay rate	Treatments (bonuses)	Available HITs
Pilot	Nov 23, 2017 (Thursday)	All day	$0.10	$0.00, $0.05, $0.10, $0.15	60
Main	Nov 26, 2017 (Sunday)	All day	$0.22	$0.00, $0.10, $0.20, $0.30	240

Design 1 Pilot Results

Name	Treatment	N	TotalTime	AvgTimePerTask	AccuracyMean	AccuracySd
Pilot 1	$0.10	54	2h 30min	5.317min	0.559	0.320
Pilot 2	$0.25	54	1h 20min	5.837min	0.673	0.246



Design 1 Pilot Analysis

$$\text{(simple) } accuracy = \theta_0 + \theta_1 * treatment$$

$$\text{(full) } accuracy = \theta_0 + \theta_1 * treatment + \theta_2 * cq1 + \theta_3 * cq2 + \theta_4 * cq3$$

Simple Model with No Covariates:

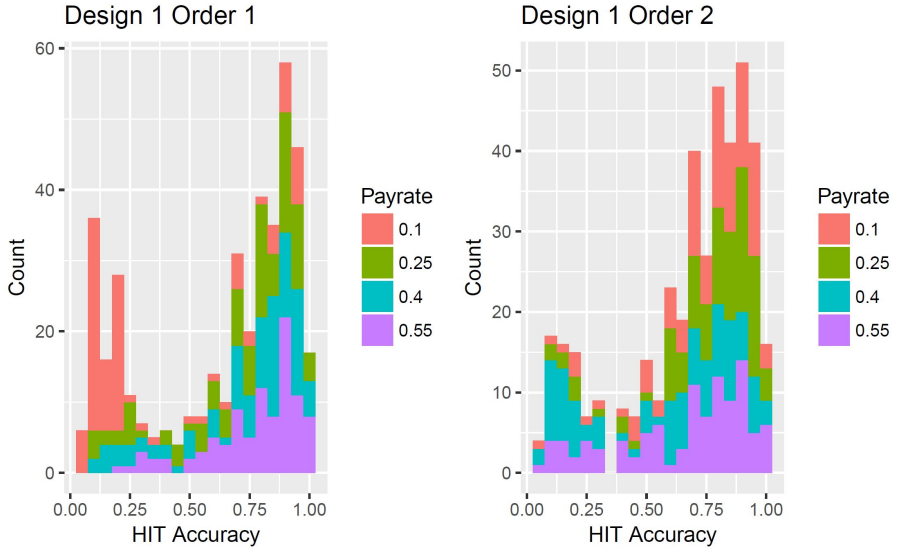
```
estimated average causal effect =  0.7555556
robust standard error =  0.3659939
95% confidence interval =  0.02993712 1.481174
p-value =  0.04142159
```

Full Model with Covariates:

```
estimated average causal effect =  0.6507679
robust standard error =  0.3467159
95% confidence interval =  -0.03702313 1.338559
p-value =  0.06341132
```

Design 1 Main Experiment Analysis

Name	Treatment	N	TotalTime	AvgTimePerTask	AccuracyMean	AccuracySd
Order 1	$0.10	102	11h 15min	0.912min	0.345	0.327
Order 1	$0.25	103	3h 15min	0.922min	0.703	0.253
Order 1	$0.40	102	0h 40min	0.866min	0.716	0.256
Order 1	$0.55	98	0h 40min	0.666min	0.774	0.192
Order 2	$0.10	102	15h 20min	0.79min	0.734	0.216
Order 2	$0.25	102	2h 40min	0.832min	0.752	0.217
Order 2	$0.40	105	2h 20min	0.855min	0.562	0.306
Order 2	$0.55	103	3h 10min	0.912min	0.656	0.265



	Dependent variable:							
	accuracy						time_spent	
	OLS						OLS	
	simple	+ covariates	+ order	+ order interaction	+ 2nd order	+ cov interaction	+ covariates	+ order interaction
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
treatment	0.288	0.288	0.286	-0.257*	0.584	0.247	-7.728	14.574***
	(0.325)	(0.318)	(0.297)	(0.103)	(0.727)	(0.264)	(12.449)	(1.220)
CQ1a lot more than half		0.130**	0.129**	0.119**	0.121**	0.174	-2.603	-2.185
		(0.042)	(0.041)	(0.038)	(0.038)	(0.121)	(4.205)	(4.326)
CQ1around half		-0.009	-0.010	-0.011	-0.008	0.008	1.847	1.850
		(0.016)	(0.015)	(0.014)	(0.014)	(0.039)	(4.467)	(4.520)
CQ2_3		-0.018	-0.020	-0.024	-0.025	0.014	-2.687	-2.545
		(0.017)	(0.015)	(0.015)	(0.015)	(0.019)	(2.002)	(2.064)
CQ3No		0.084	0.087	0.088	0.094	0.179	3.343	3.285
		(0.063)	(0.064)	(0.062)	(0.064)	(0.102)	(5.337)	(5.631)
CQ3Yes		0.168***	0.174***	0.165**	0.167**	0.232***	0.315	0.672
		(0.050)	(0.050)	(0.051)	(0.051)	(0.062)	(4.519)	(4.650)
order1			-0.049	-0.406***	-0.405***	-0.400***	-0.509	14.125***
			(0.087)	(0.119)	(0.092)	(0.115)	(3.444)	(3.574)
I(treatment2)					-1.294			
					(1.092)			
treatment:order1				1.099***	1.097***	1.089***		-45.129***
				(0.270)	(0.238)	(0.259)		(10.818)
treatment:CQ1a lot more than half						-0.170		
						(0.280)		
treatment:CQ1around half						-0.054		
						(0.092)		
treatment:CQ2_3						-0.118		
						(0.063)		
treatment:CQ3No						-0.289		
						(0.356)		
treatment:CQ3Yes						-0.213		
						(0.255)		
Constant	0.561***	0.405*	0.430**	0.624***	0.523***	0.463***	57.765***	49.804***
	(0.142)	(0.173)	(0.141)	(0.066)	(0.130)	(0.090)	(7.362)	(7.646)
Observations	817	817	817	817	817	817	817	817
R2	0.028	0.118	0.126	0.226	0.236	0.231	0.005	0.011
Adjusted R2	0.027	0.112	0.118	0.218	0.228	0.219	-0.004	0.001
Residual Std. Error	0.285 (df = 815)	0.273 (df = 810)	0.272 (df = 809)	0.256 (df = 808)	0.254 (df = 807)	0.256 (df = 803)	48.147 (df = 809)	48.028 (df = 808)
F Statistic	23.286*** (df = 1; 815)	18.118*** (df = 6; 810)	16.594*** (df = 7; 809)	29.480*** (df = 8; 808)	27.703*** (df = 9; 807)	18.553*** (df = 13; 803)	0.533 (df = 7; 809)	1.095 (df = 8; 808)

Note: *p<0.05; **p<0.01; ***p<0.001

Design 1 Main Experiment Analysis

Linear Regression [$-\infty, \infty$]

$$(3)\ aaccuracy = \theta_0 + \theta_1 * treatment + \theta_2 * cq1 + \theta_3 * cq2 + \theta_4 * cq3 + \theta_5 * order1$$

$$(4)\ accuracy = \theta_0 + \theta_1 * treatment + \theta_2 * cq1 + \theta_3 * cq2 + \theta_4 * cq3 + \theta_5 * order1 + \theta_6 * treatment * order1$$

Logistic Regression [0,1]

$$(3.1)\ logit(accuracy) = \theta_0 + \theta_1 * treatment + \theta_2 * cq1 + \theta_3 * cq2 + \theta_4 * cq3 + \theta_5 * order1$$

$$(4.1)\ logit(accuracy) = \theta_0 + \theta_1 * treatment + \theta_2 * cq1 + \theta_3 * cq2 + \theta_4 * cq3 + \theta_5 * order1 + \theta_6 * treatment * order1$$

Design 1 Models 3 & 4

Model 3: add Order indicator: ATE

estimated average causal effect = 0.2864514

Clustered standard errors = 0.2969405

.95 CI with clustered SE = [-0.2964133 0.869316]

p-value = 0.3349954

Model 4: add interaction term: ATE for Order 1

(derived using Simultaneous Tests for General Linear Hypotheses)

estimated average causal effect = 0.8423

Clustered standard errors = 0.2517

.95 CI with clustered SE = [0.348237 1.336363]

p-value = 0.000857

Model 4: add interaction term: ATE for Order 2

estimated average causal effect = -0.2566776

Clustered standard errors = 0.1028595

.95 CI with clustered SE = [-0.458581 -0.05477418]

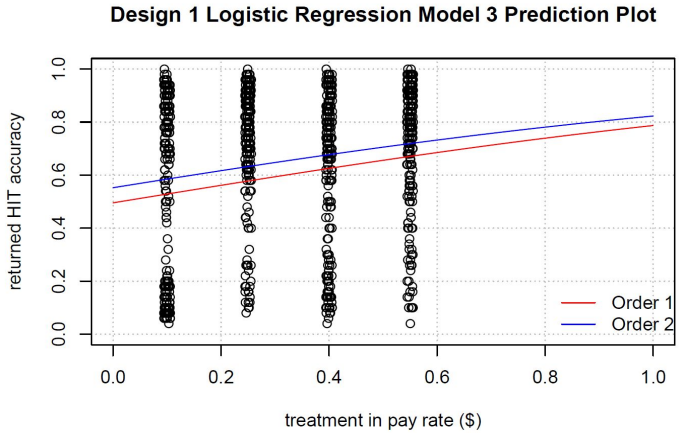
p-value = 0.01277931

Design 1 Main ATE Estimation

	Dependent variable: accuracy OLS	
	+ order (3)	+ order interaction (4)
treatment	0.286	-0.257*
	(0.297)	(0.103)
CQ1a lot more than half	0.129**	0.119**
	(0.041)	(0.038)
CQ1around half	-0.010	-0.011
	(0.015)	(0.014)
CQ2_3	-0.020	-0.024
	(0.015)	(0.015)
CQ3No	0.087	0.088
	(0.064)	(0.062)
CQ3Yes	0.174***	0.165**
	(0.050)	(0.051)
order1	-0.049	-0.406***
	(0.087)	(0.119)
I(treatment2)		
treatment:order1		1.099***
		(0.270)
Constant	0.430**	0.624***
	(0.141)	(0.066)
Observations	817	817
R2	0.126	0.226
Adjusted R2	0.118	0.218
Residual Std. Error	0.272 (df = 809)	0.256 (df = 808)
F Statistic	16.594*** (df = 7; 809)	29.480*** (df = 8; 808)

Note:

Design 1 Models 3.1 & 4.1



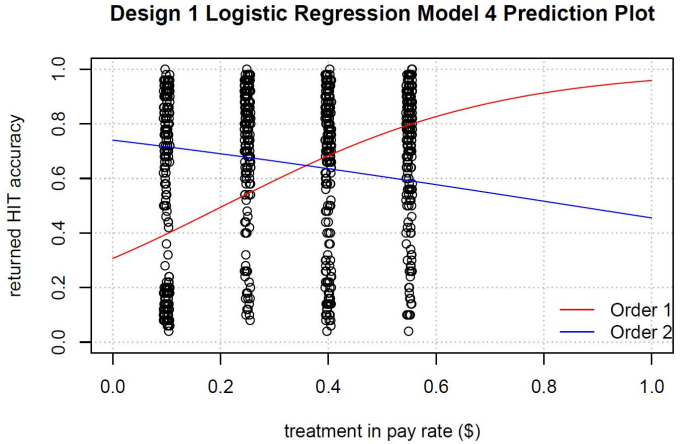
Model 3: Logistic Regression
(derived using Simultaneous Tests for General Linear Hypotheses)

estimated coefficient of treatment = 1.325

Clustered standard errors = 1.341

.95 Wald CI with clustered SE = [-1.307 3.957]

p-value = 0.323



Model 4: Logistic Regression : Order 1
(derived using Simultaneous Tests for General Linear Hypotheses)

estimated coefficient of treatment = 3.961

Clustered standard errors = 1.067

.95 Wald CI with clustered SE = [1.866581 2.936719]

p-value = 0.000205

Model 4: Logistic Regression : Order 2
(derived using Simultaneous Tests for General Linear Hypotheses)

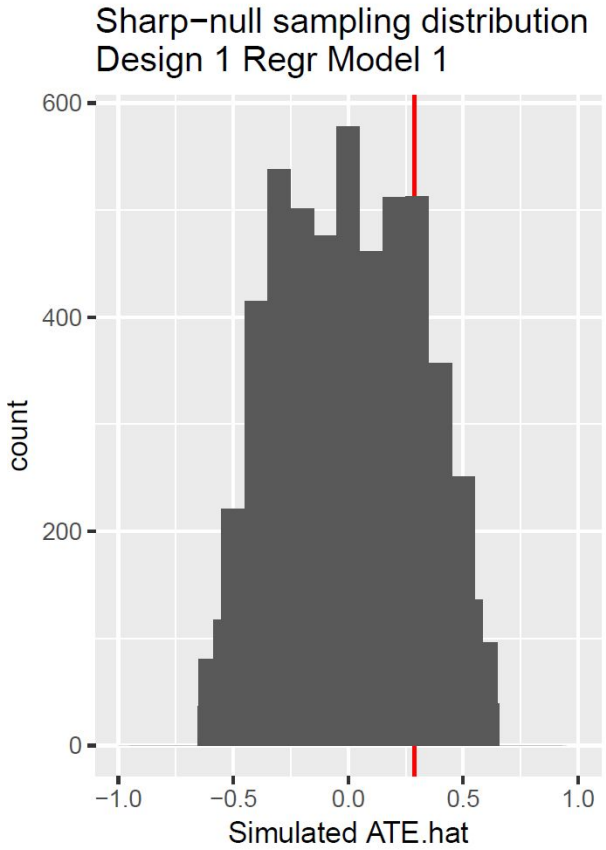
estimated coefficient of treatment = -1.226

Clustered standard errors = 0.4767

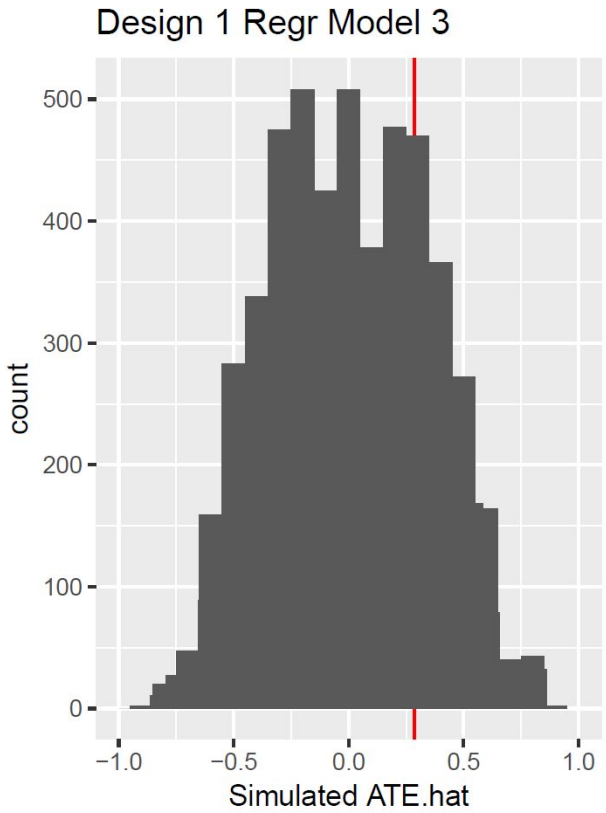
.95 Wald CI with clustered SE = [-2.161716 -0.2902835]

p-value = 0.0101

Design 1 Randomization Inference



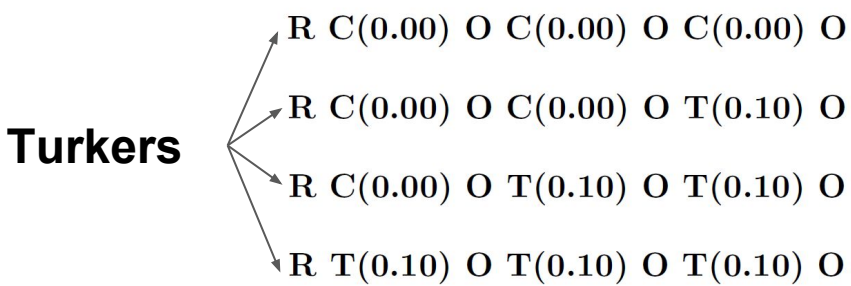
(p-val = 0.411)



(p-val = 0.477)

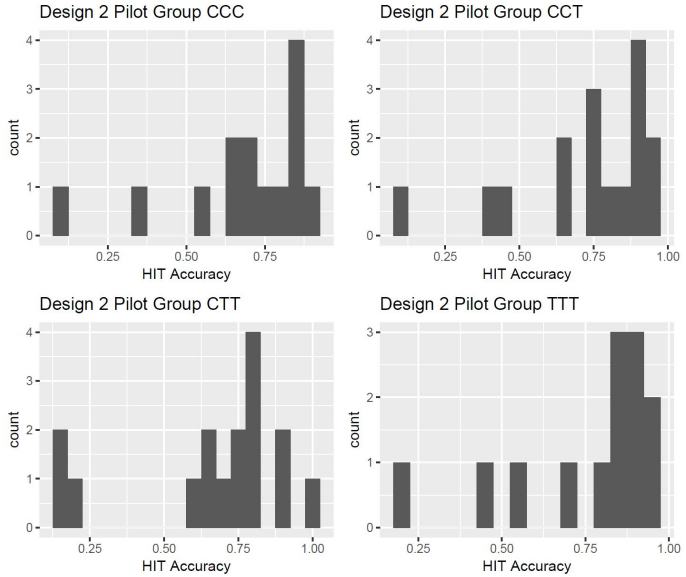
- Group CCC : Turkers in this group do not receive bonuses in any sessions
- Group CCT : Turkers in this group only receive bonuses in the third session
- Group CTT : Turkers in this group only receive bonuses in the second and third sessions
- Group TTT : Turkers in this group receive bonsues in all three sessions

Design 2 Notation: Stepped-Wedge Design



Design 2 Pilot Results

Group	Treatment	N	TotalTime	AvgTimePerTask	AccuracyMean	AccuracySd
All	/	59	20h 55min	6.651min	0.707	0.235
~~CCC~~	~~0.00~~	~~14~~	/	7.685min	0.681	0.217
CCT	$0.05	16	/	5.549min	0.729	0.236
CTT	$0.10	16	/	8.199min	0.663	0.27
TTT	$0.15	13	/	4.989min	0.763	0.219
Attritiers		12	/	/	/	/



All covariate checks are passed

Covariate Balance Check Design 2 Pilot

	Dependent variable:						
	CQ1.1 (1)	CQ1.2 (2)	CQ1.3 (3)	CQ2 (4)	CQ3.1 (5)	CQ3.2 (6)	CQ3.3 (7)
groupCCT	−0.089	0.063	0.027	0.259	0.223	−0.152	−0.071
	(0.147)	(0.195)	(0.179)	(0.310)	(0.145)	(0.145)	(0.145)
groupCTT	0.098	0.063	−0.161	−0.366	−0.027	−0.089	0.116
	(0.171)	(0.195)	(0.157)	(0.260)	(0.179)	(0.179)	(0.179)
groupTTT	−0.137	0.038	0.099	0.379	0.055	−0.214	0.159
	(0.143)	(0.208)	(0.196)	(0.319)	(0.181)	(0.181)	(0.181)
Constant	0.214	0.500***	0.286*	1.929***	0.714***	0.214	0.071
	(0.118)	(0.144)	(0.130)	(0.202)	(0.130)	(0.130)	(0.130)
Observations	59	59	59	59	59	59	59
R2	0.054	0.003	0.046	0.126	0.059	0.064	0.080
Adjusted R2	0.002	−0.052	−0.006	0.078	0.007	0.013	0.030
Residual Std. Error (df = 55)	0.392	0.515	0.450	0.787	0.416	0.303	0.321
F Statistic (df = 3; 55)	1.042	0.048	0.889	2.642	1.145	1.254	1.602

	Dependent variable:							
	CQ4.1 (1)	CQ4.2 (2)	CQ4.3 (3)	CQ4.4 (4)	CQ4.5 (5)	CQ5.1 (6)	CQ5.2 (7)	CQ5.3 (8)
groupCCT	−0.152	0.063	0.036	0.107	−0.054	−0.009	0.009	0.000
	(0.135)	(0.065)	(0.165)	(0.153)	(0.192)	(0.194)	(0.194)	(0.000)
groupCTT	−0.152	0.188	0.098	−0.018	−0.116	−0.009	0.009	0.000
	(0.135)	(0.104)	(0.171)	(0.134)	(0.189)	(0.194)	(0.194)	(0.000)
groupTTT	−0.137	0.154	0.093	−0.143	0.033	−0.110	0.033	0.077
	(0.143)	(0.108)	(0.182)	(0.101)	(0.207)	(0.207)	(0.207)	(0.080)
Constant	0.214	−0.000	0.214	0.143	0.429**	0.571***	0.429**	−0.000
	(0.118)		(0.118)	(0.101)	(0.142)	(0.142)	(0.142)	(0.000)
Observations	59	59	59	59	59	59	59	59
R2	0.044	0.060	0.008	0.065	0.013	0.007	0.001	0.061
Adjusted R2	−0.009	0.009	−0.046	0.014	−0.040	−0.047	−0.054	0.010
Residual Std. Error (df = 55)	0.306	0.304	0.459	0.343	0.502	0.514	0.514	0.130
F Statistic (df = 3; 55)	0.834	1.167	0.154	1.279	0.248	0.138	0.010	1.191

Note: *p<0.05; **p<0.01; ***p<0.001

Design 2 Pilot Analysis

(simple) $overall\ accuracy = \theta_0 + \theta_1 * treatment$

(full) $overall\ accuracy = \theta_0 + \theta_1 * treatment + \theta_2 * cq1 + \theta_3 * cq2 + \theta_4 * cq3 + \theta_5 * cq4 + \theta_6 * cq5$

(simple) $total\ timespent = \theta_0 + \theta_1 * treatment$

(full) $total\ timespent = \theta_0 + \theta_1 * treatment + \theta_2 * cq1 + \theta_3 * cq2 + \theta_4 * cq3 + \theta_5 * cq4 + \theta_6 * cq5$

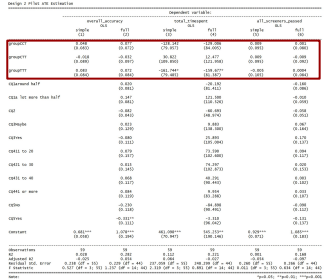
(simple) $all\ screeners\ passed = \theta_0 + \theta_1 * treatment$

(full) $all\ screeners\ passed = \theta_0 + \theta_1 * treatment + \theta_2 * cq1 + \theta_3 * cq2 + \theta_4 * cq3 + \theta_5 * cq4 + \theta_6 * cq5$

```
Design 2 Pilot, Simple Model : Effect of being in group CCT rather than CCC on accuracy

estimated average causal effect =  0.0484944
robust standard error =  0.08287031
95% confidence interval =  -0.1175814 0.2145702
p-value =  0.5608172
```

```
Design 2 Pilot, Simple Model : Effect of being in group CTT rather than CCC on accuracy

estimated average causal effect =  -0.01768207
robust standard error =  0.08916341
95% confidence interval =  -0.1963695 0.1610054
p-value =  0.8435331
```

```
Design 2 Pilot, Simple Model : Effect of being in group TTT rather than CCC on accuracy

estimated average causal effect =  0.08252532
robust standard error =  0.08371201
95% confidence interval =  -0.0852373 0.2502879
p-value =  0.3285365
```



Design 2 Main Experiment Analysis

Group	Treatment	N	TotalTime	AvgTimePerTask	AccuracyMean	AccuracySd
All	/	243	10h 2min	6.239min	0.807	0.197
CCC	0.00	61	/	5.854min	0.823	0.179
CCT	$0.05	58	/	6.22min	0.821	0.205
CTT	$0.10	60	/	5.834min	0.783	0.199
TTT	$0.15	64	/	7.004min	0.8	0.206
Attritiers		49	/	/	/	/

Differential Attrition Check Design 2 Main

	Dependent variable:		
	attrited		
	(1)	(2)	(3)
groupCCT	0.031		-0.552
	(0.060)		(0.592)
groupCTT	0.014		-0.504
	(0.059)		(0.340)
groupTTT	-0.030		-0.018
	(0.059)		(0.359)
CQ1around half		-0.070	-0.305*
		(0.063)	(0.133)
CQ1a lot more than half		-0.133*	-0.446***
		(0.064)	(0.130)
groupCCT:CQ1around half			0.405
			(0.215)
groupCTT:CQ1around half			0.257
			(0.186)
groupTTT:CQ1around half			0.282
			(0.176)
groupCCT:CQ1a lot more than half			0.643**
			(0.208)
groupCTT:CQ1a lot more than half			0.204
			(0.179)
groupTTT:CQ1a lot more than half			0.396*
			(0.178)

Design 2 Main Analysis Approach 1

Linear -- accuracy

(1) $overall\ accuracy = \theta_0 + \theta_1 * treatment$

(2) $overall\ accuracy = \theta_0 + \theta_1 * treatment + \theta_2 * cq1 + \theta_3 * cq2 + \theta_4 * cq3 + \theta_5 * cq4 + \theta_6 * cq5$

Logistic -- accuracy

(3) $logit(overall\ accuracy) = \theta_0 + \theta_1 * treatment$

(4) $logit(overall\ accuracy) = \theta_0 + \theta_1 * treatment + \theta_2 * cq1 + \theta_3 * cq2 + \theta_4 * cq3 + \theta_5 * cq4 + \theta_6 * cq5$

Linear time-spent

(5) $total\ time\ spent = \theta_0 + \theta_1 * treatment$

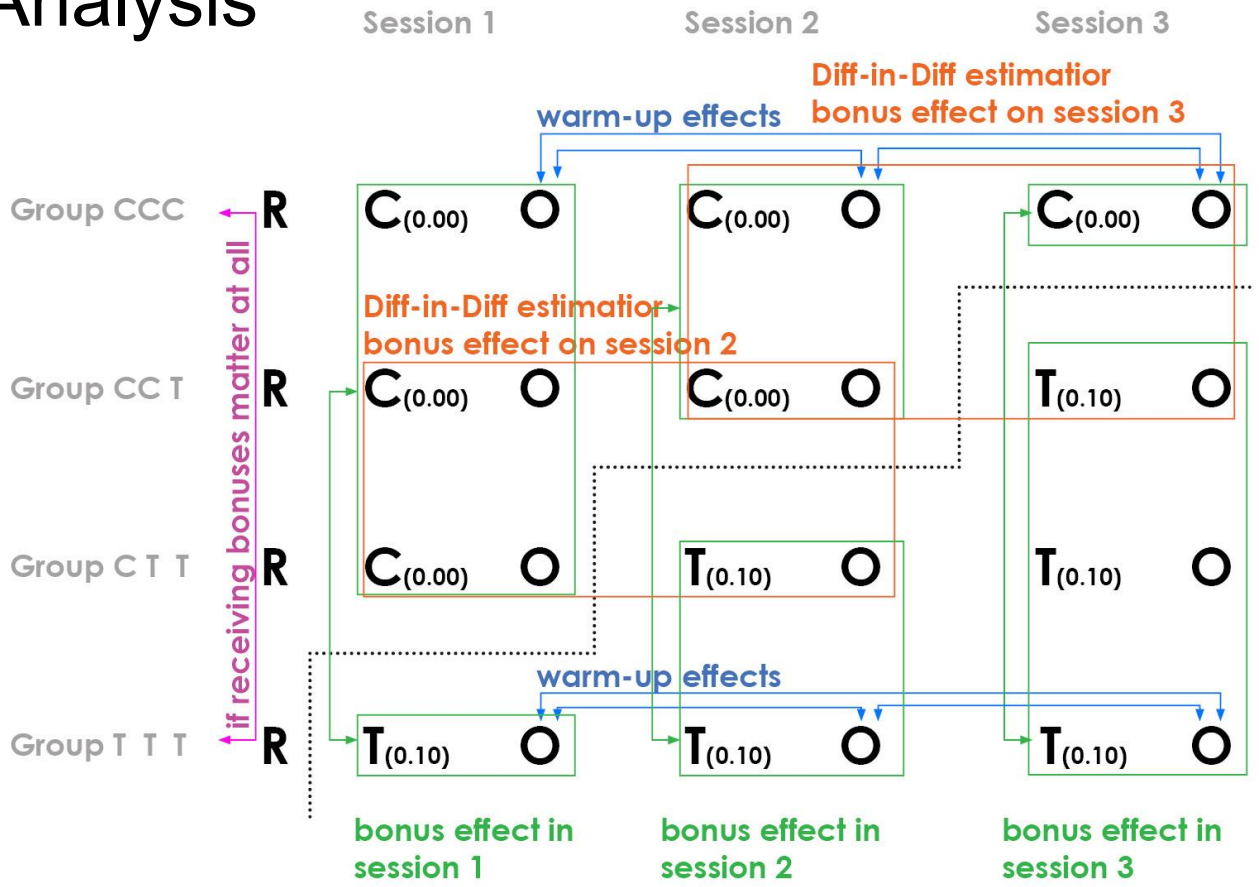
(6) $total\ time\ spent = \theta_0 + \theta_1 * treatment + \theta_2 * cq1 + \theta_3 * cq2 + \theta_4 * cq3 + \theta_5 * cq4 + \theta_6 * cq5$

Design 2 Main experiment ATE Estimation

	Dependent variable:					
	overall_accuracy				total_timespent	
	OLS		logistic		OLS	
	simple (1)	full (2)	simple (3)	full (4)	simple (5)	full (6)
groupCCT	-0.001	0.013	-0.010	0.084	21.946	24.713
	(0.036)	(0.033)	(0.243)	(0.242)	(31.790)	(34.618)
groupCTT	-0.039	-0.033	-0.249	-0.228	-1.232	-5.639
	(0.035)	(0.032)	(0.220)	(0.208)	(28.794)	(28.817)
groupTTT	-0.023	-0.002	-0.150	0.001	68.965	68.494
	(0.035)	(0.030)	(0.226)	(0.210)	(37.967)	(38.113)
CQ1around half		0.081		0.470		-48.134
		(0.047)		(0.255)		(33.606)
CQ1a lot more than half		0.115*		0.718**		-30.540
		(0.047)		(0.263)		(31.103)
CQ2		0.024		0.168		15.220
		(0.016)		(0.115)		(15.956)
CQ3Maybe		-0.153		-0.708		-141.388
		(0.087)		(0.406)		(82.260)
CQ3Yes		0.101*		0.602*		-100.141
		(0.050)		(0.267)		(53.943)
CQ411 to 20		-0.065		-0.445		-14.201
		(0.055)		(0.369)		(60.593)
CQ421 to 30		-0.030		-0.239		0.953
		(0.045)		(0.336)		(54.521)
CQ431 to 40		0.008		0.039		19.894
		(0.049)		(0.360)		(60.118)
CQ441 or more		-0.032		-0.265		-51.929
		(0.041)		(0.298)		(47.400)
CQ5No		0.181*		1.179**		161.226***
		(0.080)		(0.409)		(42.751)
CQ5Yes		0.114		0.719		156.301***
		(0.082)		(0.414)		(43.122)
Constant	0.823***	0.482***	1.534***	-0.495	351.246***	323.026***
	(0.023)	(0.097)	(0.158)	(0.476)	(22.820)	(64.328)
Observations	243	243	243	243	243	243
R2	0.007	0.213			0.024	0.079
Adjusted R2	-0.006	0.165			0.012	0.022
Log Likelihood			-83.451	-70.854		
Akaike Inf. Crit.			174.903	171.707		
Residual Std. Error	0.197 (df = 239)	0.180 (df = 228)			185.162 (df = 239)	184.203 (df = 228)
F Statistic	0.543 (df = 3; 239)	4.415*** (df = 14; 228)			1.972 (df = 3; 239)	1.391 (df = 14; 228)

Note: *p<0.05; **p<0.01; ***p<0.001

Design 2 Main Analysis Approach 2



Design 2 Main Analysis Approach 2

(7) $accuracy = \beta_{intercept} + \beta_{GS} group * session$ (Compressed Form)

(7.1) $accuracy = \beta_1 + \beta_2 * groupCCT + \beta_3 * groupCTT + \beta_4 * groupTTT + \beta_5 * session2 + \beta_6 * session3 + \beta_7 groupCCT : session2 + \beta_8 groupCTT : session2 + \beta_9 groupTTT : session2 + \beta_{10} groupCCT * session3 + \beta_{11} groupCTT * session3 + \beta_{12} groupTTT * session3$ (Expanded Form)

Notation	Meaning	Linear Combination
CCC S2-S1	effect of being in round two rather than round one in group CCC	b_5
CCC S3-S1	effect of being in round three rather than round one in group CCC	b_6
CCC S3-S2	effect of being in round three rather than round two in group CCC	b_6 - b_5
CCT S2-S1	effect of being in round two rather than round one in group CCT	b_5 + b_7
CCT S3-S1	effect of being in round three rather than round one in group CCT	b_6 + b_10
CCT S3-S2	effect of being in round three rather than round two in group CCT	b_6 - b_5 + b_10 - b_7
CTT S2-S1	effect of being in round two rather than round one in group CTT	b_5 + b_8
CTT S3-S1	effect of being in round three rather than round one in group CTT	b_6 + b_11
CTT S3-S2	effect of being in round three rather than round two in group CTT	b_6 - b_5 + b_11 - b_8
TTT S2-S1	effect of being in round two rather than round one in group TTT	b_5 + b_9
TTT S3-S1	effect of being in round three rather than round one in group TTT	b_6 + b_12
TTT S3-S2	effect of being in round three rather than round two in group TTT	b_6 - b_5 + b_12 - b_9
T-C S1	effect of receiving treatment in Round one	b_4 - (b_3 + b_2)/3
T-C S2	effect of receiving treatment in Round two (regardless of round one status)	(-b_2 + b_3 + b_4 - b_7 + b_8 + b_9)/2
T-C S3	effect of receiving treatment in Round three (regardless of round one or two status)	(b_4 + b_12) - (b_2 + b_3 + b_10 + b_11)/3
TTT-CCC	effect of being in a group TTT that receive all three rounds of treatment rather than all control group CCC	b_4 + (b_9 + b_12)/3
(T)-CCC	effect of being in any treatment groups rather than all control group CCC	(b_2 + b_3 + b_4)/3 + (b_7 + b_8 + b_9 + b_10 + b_11 + b_12)/9
D_in_D S3	effect on Round 3, of Receiving Treatment in Round 3 but not earlier	b_10 - b_7
D_in_D S2	effect on Round 2, of Receiving Treatment in Round 2 but not earlier	- (1/2)*b_7 + b_8

Design 2 Main Analysis Approach 2

Simultaneous Tests for General Linear Hypotheses

Fit: lm(formula = round_accuracy ~ group * round, data = by_Session.

Linear Hypotheses:

	Estimate	Std. Error	t value	Pr(>|t|)
CCC S2-S1 == 0	-0.0173578	0.0369009	-0.470	1.000
CCC S3-S1 == 0	0.0144648	0.0355561	0.407	1.000
CCC S3-S2 == 0	0.0318226	0.0321593	0.990	0.988
CCT S2-S1 == 0	-0.0223124	0.0402194	-0.555	1.000
CCT S3-S1 == 0	0.0182556	0.0398488	0.458	1.000
CCT S3-S2 == 0	0.0405680	0.0431454	0.940	0.992
CTT S2-S1 == 0	-0.0009804	0.0389598	-0.025	1.000
CTT S3-S1 == 0	0.0098039	0.0393088	0.249	1.000
CTT S3-S2 == 0	0.0107843	0.0379124	0.284	1.000
TTT S2-S1 == 0	-0.0101103	0.0389303	-0.260	1.000
TTT S3-S1 == 0	0.0238971	0.0385949	0.619	1.000
TTT S3-S2 == 0	0.0340074	0.0392051	0.867	0.996
T-C S1 == 0	-0.0137755	0.0314239	-0.438	1.000
T-C S2 == 0	-0.0210181	0.0273768	-0.768	0.999
T-C S3 == 0	-0.0040532	0.0314754	-0.129	1.000
TTT-CCC == 0	-0.0229327	0.0213467	-1.074	0.979
(T)-CCC == 0	-0.0211889	0.0170422	-1.243	0.943
D_in_D S3 == 0	0.0087454	0.0538121	0.163	1.000
D_in_D S2 == 0	0.0188547	0.0475677	0.396	1.000

(Adjusted p values reported -- single-step method)

Rboust Standard Errors are applied

Design 2 Main Analysis Approach 3

- **Y_00** (untreated during preceding and current session)
- **Y_01** (untreated during preceding session but treated during current session)
- **Y_11** (treated during preceding and current session)

```
[1] "Treatment Condition Table:"
```

V1	worker_id	one	two	three
1	A10HEV2856GIQL	01	11	11
2	A10TT0QTE6FNQQ	01	11	11
3	A10XK0DWEXO5BY	00	01	11
4	A1131ZL7O5GM6R	00	01	11
5	A11QDNT3W7DT7K	00	00	01

```
[1] "Observed Outcomes Table:"
```

V1	worker_id	one	two	three
1	A10HEV2856GIQL	0.4705882	0.5294118	0.1176471
2	A10TT0QTE6FNQQ	1.0000000	0.9411765	1.0000000
3	A10XK0DWEXO5BY	0.4705882	0.3529412	0.6470588
4	A1131ZL7O5GM6R	1.0000000	0.9411765	0.9411765
5	A11QDNT3W7DT7K	0.8235294	0.7058824	0.8235294

Design 2 Main Analysis Approach 3

Treatment_Condition	Week_1	Week_2	Week_3
Pr(00)	0.75	0.50	0.25
Pr(01)	0.25	0.25	0.25
Pr(11)	0.00	0.25	0.50

$$\hat{E}[Y_{01} - Y_{00}] = \frac{\frac{\sum_{S1} Y_{01}}{0.25} + \frac{\sum_{S2} Y_{01}}{0.25} + \frac{\sum_{S3} Y_{01}}{0.25}}{\frac{64}{0.25} + \frac{60}{0.25} + \frac{60}{0.25}} - \frac{\frac{\sum_{S1} Y_{00}}{0.75} + \frac{\sum_{S2} Y_{00}}{0.50} + \frac{\sum_{S3} Y_{00}}{0.25}}{\frac{179}{0.75} + \frac{119}{0.50} + \frac{61}{0.25}}$$

```
##
## Estimated Combined Immediate Effect: -0.02847309
```

$$\hat{E}[Y_{11} - Y_{00}] = \frac{\frac{\sum_{S2} Y_{11}}{0.25} + \frac{\sum_{S3} Y_{11}}{0.50}}{\frac{64}{0.25} + \frac{124}{0.50}} - \frac{\frac{\sum_{S2} Y_{00}}{0.50} + \frac{\sum_{S3} Y_{00}}{0.25} +}{\frac{119}{0.50} + \frac{61}{0.25}}$$

```
##
## Estimated Combined Immediate and Lagged Effect: -0.02602659
```

Design 2 Main Analysis Approach 3

Observed Y00 Table:

V1	worker_id	one	two	three
1	A10HEV2856GIQL	NA	NA	NA
2	A10TT0QTE6FNQQ	NA	NA	NA
3	A10XK0DWEXO5BY	0.4705882	NA	NA
4	A1131ZL7O5GM6R	1.0000000	NA	NA
5	A11QDNT3W7DT7K	0.8235294	0.7058824	NA

Hypothetical Y01 Table:

V1	worker_id	one	two	three
1	A10HEV2856GIQL	0.4990613	0.5554384	0.1436737
2	A10TT0QTE6FNQQ	1.0284731	0.9672031	1.0260266
3	A10XK0DWEXO5BY	0.4705882	0.3814143	0.6730854
4	A1131ZL7O5GM6R	1.0000000	0.9696496	0.9672031
5	A11QDNT3W7DT7K	0.8235294	0.7058824	0.8520025

Observed Y01 Table:

V1	worker_id	one	two	three
1	A10HEV2856GIQL	0.4705882	NA	NA
2	A10TT0QTE6FNQQ	1.0000000	NA	NA
3	A10XK0DWEXO5BY	NA	0.3529412	NA
4	A1131ZL7O5GM6R	NA	0.9411765	NA
5	A11QDNT3W7DT7K	NA	NA	0.8235294

Hypothetical Y01 Table:

V1	worker_id	one	two	three
1	A10HEV2856GIQL	0.4705882	0.5269653	0.1152006
2	A10TT0QTE6FNQQ	1.0000000	0.9387300	0.9975535
3	A10XK0DWEXO5BY	0.4421151	0.3529412	0.6446123
4	A1131ZL7O5GM6R	0.9715269	0.9411765	0.9387300
5	A11QDNT3W7DT7K	0.7950563	0.6774093	0.8235294

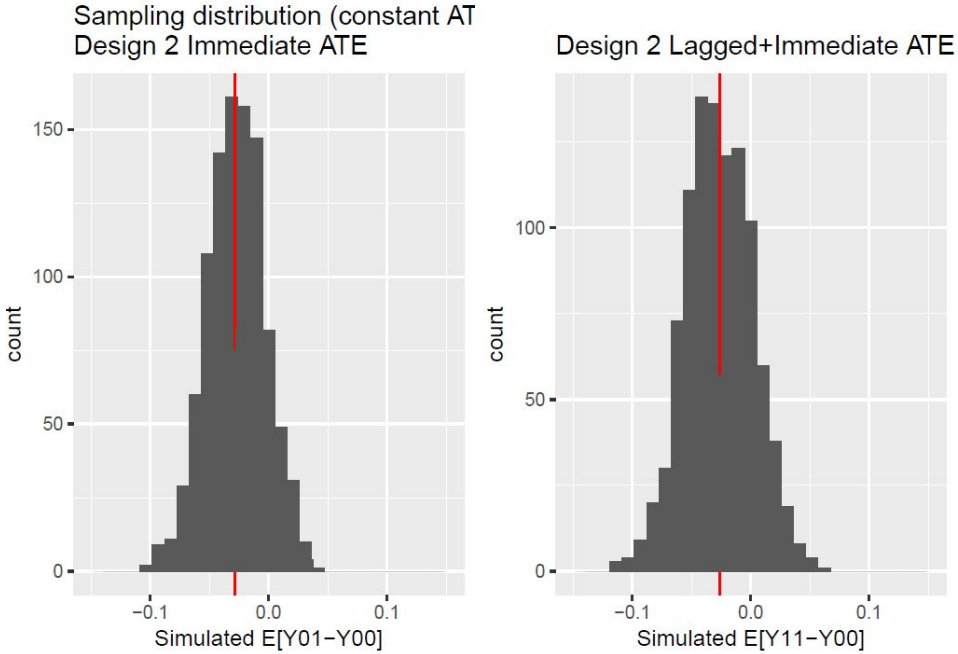
Observed Y11 Table:

V1	worker_id	one	two	three
1	A10HEV2856GIQL	NA	0.5294118	0.1176471
2	A10TT0QTE6FNQQ	NA	0.9411765	1.0000000
3	A10XK0DWEXO5BY	NA	NA	0.6470588
4	A1131ZL7O5GM6R	NA	NA	0.9411765
5	A11QDNT3W7DT7K	NA	NA	NA

Hypothetical Y01 Table:

V1	worker_id	one	two	three
1	A10HEV2856GIQL	0.4730347	0.5294118	0.1176471
2	A10TT0QTE6FNQQ	1.0024465	0.9411765	1.0000000
3	A10XK0DWEXO5BY	0.4445616	0.3553877	0.6470588
4	A1131ZL7O5GM6R	0.9739734	0.9436230	0.9411765
5	A11QDNT3W7DT7K	0.7975028	0.6798558	0.8259759

Design 2 Main Analysis Approach 3



Estimated Confidence Interval of E[Y01-Y00]--immediate treatment effect:

2.5%	97.5%
-0.07564641	0.02009593

Estimated Confidence Interval of E[Y11-Y00]--combined immediate and lagged treatment effect:

2.5%	97.5%
-0.08192224	0.03075000

Summary of All Results

Design 1 Notation: Between Subject Design

- T-tests
- Linear, Logistic Regression Models
- Randomization Inference

- Significance only observed for HTE regarding the two publish orders on two weekends

Design 2 Notation: Stepped-Wedge Design

- Between group comparisons
- Design notation comparisons
- Immediate and Lag Effects

- Neither statistical nor practical significance is observed in any case

Within subjects design limitations

10 cent treatment accuracy & time to completion varied between order 1 & 2

Several possible reasons

- Exhausting turkers
- Timing
- Ordering

Simple fix: add at least two more orderings.

Future proposal for differential attrition

More regular exposures to dogs = less likely to attrit w/ no bonuses assigned. But the opposite happens when bonuses are assigned.

To try to mitigate the bias, re-invite the same attrited turkers to complete the task again, under the same conditions.

How our findings generalize

Design 1, order 1 was completed over Veteran’s day weekend.

Design 2 has ‘bonus’, i.e. not displayed upfront, payments.

Increasing upper end of payments, (e.g. $0.80) may cause differences in accuracy at the $0.55 previously upper end level.

It’s difficult to generalize the AMT market to the labor market at large, and the dog breed identification task may not even generalize to more serious tasks.


The End

Table of Contents - to be removed

1. Introduction and Background
2. Research Hypothesis and Experimental Design
3. Design 1 Study
    a. Schedule
    b. Pilot Study
    c. Data
    d. Analysis
4. Design 2 Study
    a. Schedule
    b. Pilot Study
    c. Data
    d. Analysis
5. Overall Results
6. Limitations and Future Development

We investigate the relationship between reward and productivity.

- Does paying more for a job lead to higher quality results?
- AMT and Qualtrics
- Two experimental designs:
  - Design 1: Between-subjects experiment
  - Design 2: Stepped-wedge experiment
- Prior Studies

We hypothesize that **higher reward per HIT(human intelligence task) would on average lead to higher returned task accuracy, or turker performance**

- Treatment: pay rate in $ ($0.10 - $0.55/HIT)
- Outcome: % of dog image classification questions
- Multiple-choice questions
- Covariate questions:
  - Familiarity with dogs
  - Preference for media
  - AMT experience
- Data: Stanford Dogs Dataset

Experimental Design

- Problem: Self-selection and biased estimates
- Design 1: Between-subject comparison
  - Unit of analysis: HIT
  - Aim for higher accuracy
  - Selection bias and attrition not a concern
- Design 2: Stepped-wedge comparison
  - Unit of analysis: worker
  - Aim to measure motivation
  - Selection bias and attrition are huge concerns

(Add design notations)

Design 1: Between-Subjects

Publish Order	Date	Time Frame	Treatment (Pay Rate)	Available HITs
Pilot	Oct 28, 2017 (Saturday)	Morning	$0.10	50
Pilot	Oct 29, 2017 (Sunday)	Afternoon	$0.25	50
1	Nov 11, 2017 (Saturday)	Morning	$0.10	100
1	Nov 11, 2017 (Saturday)	Afternoon	$0.55	100
1	Nov 12, 2017 (Sunday)	Morning	$0.25	100
1	Nov 12, 2017 (Sunday)	Afternoon	$0.40	100
2	Nov 18, 2017 (Saturday)	Morning	$0.40	100
2	Nov 18, 2017 (Saturday)	Afternoon	$0.25	100
2	Nov 19, 2017 (Sunday)	Morning	$0.55	100
2	Nov 19, 2017 (Sunday)	Afternoon	$0.10	100

Design 1 Pilot

Name	Treatment	N	TotalTime	AvgTimePerTask	AccuracyMean	AccuracySd
Pilot 1	$0.10	54	2h 30min	5.317min	0.559	0.320
Pilot 2	$0.25	54	1h 20min	5.837min	0.673	0.246

- Bimodal distribution of accuracy
- Models used:
-

$$\text{(simple) } accuracy = \theta_0 + \theta_1 * treatment$$

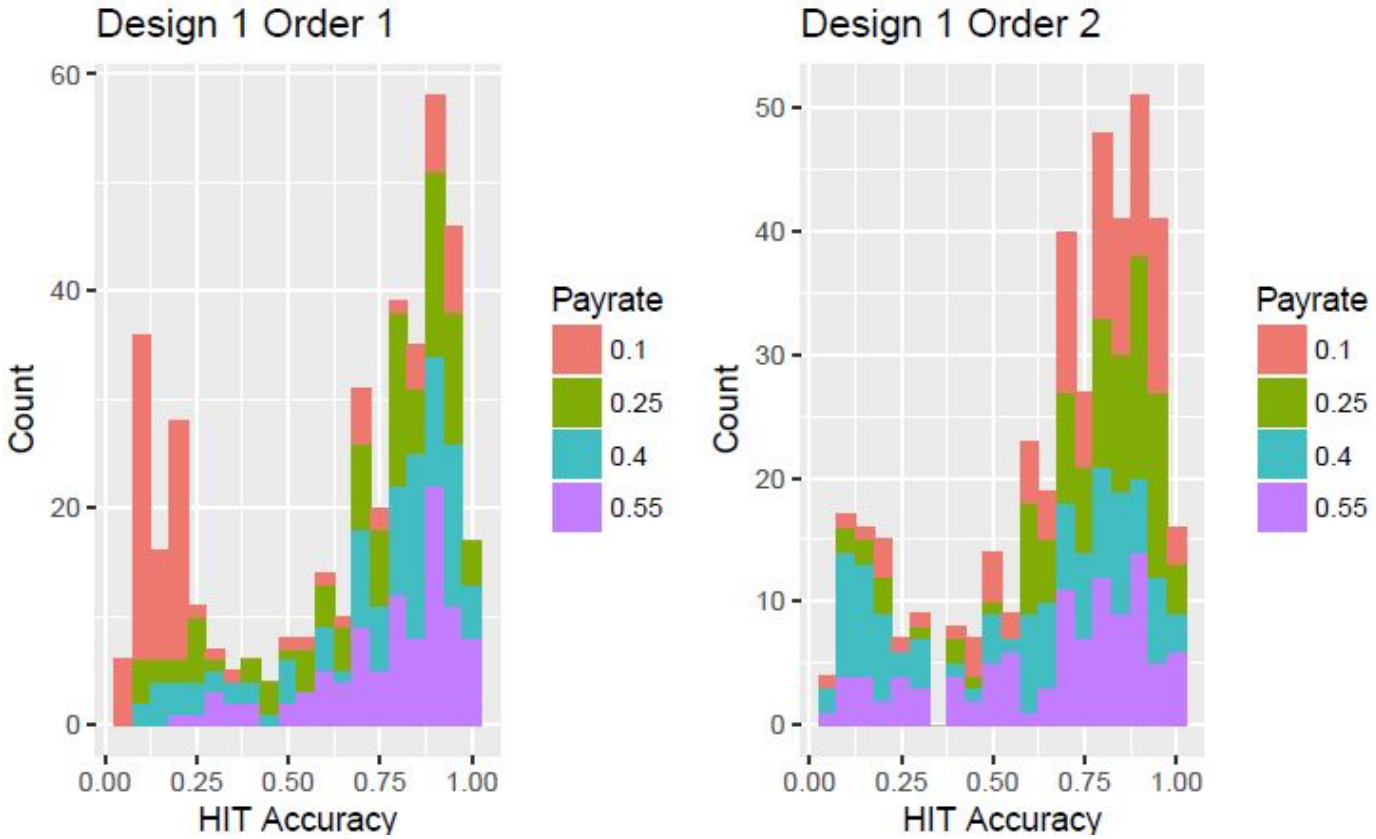
$$\text{(full) } accuracy = \theta_0 + \theta_1 * treatment + \theta_2 * cq1 + \theta_3 * cq2 + \theta_4 * cq3$$

	Dependent variable:	
	accuracy	
	simple	full
	(1)	(2)
treatment	0.756*	0.651
	(0.369)	(0.359)
CQ1a lot more than half		0.146
		(0.081)
CQ1around half		0.002
		(0.082)
CQ2_3		-0.072*
		(0.036)
CQ3No		0.119
		(0.146)
CQ3Yes		0.126
		(0.127)
Constant	0.484***	0.472**
	(0.077)	(0.168)
Observations	108	108
R2	0.039	0.154
Adjusted R2	0.030	0.103
Residual Std. Error	0.285 (df = 106)	0.274 (df = 101)
F Statistic	4.262* (df = 1; 106)	3.054** (df = 6; 101)
Note:		*p<0.05; **p<0.01; ***p<0.001

Design 1 Main Experiment

Name	Treatment	N	TotalTime	AvgTimePerTask	AccuracyMean	AccuracySd
Order 1	$0.10	102	11h 15min	0.912min	0.345	0.327
Order 1	$0.25	103	3h 15min	0.922min	0.703	0.253
Order 1	$0.40	102	0h 40min	0.866min	0.716	0.256
Order 1	$0.55	98	0h 40min	0.666min	0.774	0.192
Order 2	$0.10	102	15h 20min	0.79min	0.734	0.216
Order 2	$0.25	102	2h 40min	0.832min	0.752	0.217
Order 2	$0.40	105	2h 20min	0.855min	0.562	0.306
Order 2	$0.55	103	3h 10min	0.912min	0.656	0.265

Design 1 Accuracy Distribution



Design 1 Analysis

- Linear Regression:

$$(3)\ aaccuracy = \theta_0 + \theta_1 * treatment + \theta_2 * cq1 + \theta_3 * cq2 + \theta_4 * cq3 + \theta_5 * order1$$

$$(4)\ accuracy = \theta_0 + \theta_1 * treatment + \theta_2 * cq1 + \theta_3 * cq2 + \theta_4 * cq3 + \theta_5 * order1 + \theta_6 * treatment * order1$$

- Logistic Regression:
  - Accuracy outcome is bounded by [0,1]

$$(3.1)\ logit(accuracy) = \theta_0 + \theta_1 * treatment + \theta_2 * cq1 + \theta_3 * cq2 + \theta_4 * cq3 + \theta_5 * order1$$

$$(4.1)\ logit(accuracy) = \theta_0 + \theta_1 * treatment + \theta_2 * cq1 + \theta_3 * cq2 + \theta_4 * cq3 + \theta_5 * order1 + \theta_6 * treatment * order1$$

- Adding LR Interprettion in important here

	Dependent variable:	
	accuracy	
	logistic	
	covariates + order	+ interaction
	(1)	(2)
treatment	1.325	-1.226*
	(1.341)	(0.477)
CQ1a lot more than half	0.619***	0.594***
	(0.161)	(0.164)
CQ1around half	-0.047	-0.048
	(0.065)	(0.062)
CQ2_3	-0.094	-0.110
	(0.069)	(0.071)
CQ3No	0.358	0.383
	(0.274)	(0.276)
CQ3Yes	0.742***	0.726**
	(0.221)	(0.232)
order1	-0.228	-1.862***
	(0.396)	(0.488)
treatment:order1		5.187***
		(1.167)
Constant	-0.316	0.565
	(0.618)	(0.299)
Observations	817	817
Log Likelihood	-453.580	-423.436
Akaike Inf. Crit.	923.161	864.873
Note:	*p<0.05; **p<0.01; ***p<0.001	